Vastamediauutisten laskennallinen kehysanalyysi

Pihla Toivanen

Pro gradu
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Helsinki, 8. huhtikuuta 2019

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen laitos

Tekijä — Författare — Author
Pihla Toivanen

Työn nimi — Arbetets titel — Title

Vastamediauutisten laskennallinen kehysanalyysi

Oppiaine — Läroämne — Subject
Datatiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu	8. huhtikuuta 2019	57

Tiivistelmä — Referat — Abstract

Valeuutiset ovat viime vuosina nousseet merkittäväksi yhteiskunnallisen keskustelun aiheeksi niin Suomessa kuin ulkomaillakin. Esimerkiksi vuoden 2016 yhdysvaltojen presidentinvaalien aikana jotkin valeuutiset levisivät laajemmalle kuin suosituimmat valtamediauutiset, ja valeuutisten onkin arveltu vaikuttaneen merkittävästi Trumpin voittoon kyseisissä vaaleissa.

Aiemmasta suomalaisesta tutkimuksesta tiedetään, että Suomessa valeuutiset eivät aina sisällä suoraan virheellistä tietoa, ja tämän vuoksi suomalaisia valemedioita kutsutaan myös vastamedioiksi. Tiedetään myös, että suomalaisissa vastamediauutisissa kehystetään usein valtamedian uutisia tukemaan vastamedian omaa agendaa.

Kehystämisellä tarkoitetaan viestinnän tutkimuksessa prosessia, jolla valikoinnin, poissulkemisen ja esimerkiksi metaforien ja iskulauseiden avulla muokataan mediaesityksen tulkintaa. Kehyksen käsite sekä kehysanalyysi ovat saaneet alkunsa sosiaalipsykologiasta ja levinneet sittemmin mediatutkimukseen. Laskennallisesti kehysanalyysiä on tehty sekä ohjatuilla että ohjaamattomilla koneoppimismenetelmillä, mutta yksikään näistä menetelmistä ei ole vakiintunut kehyksen operationalisoinnin monikäsitteisyyden vuoksi.

Tämän tutkielman tarkoituksena on selvittää, millaisilla prosesseilla suomalainen vastamedia uudelleenkehystää valtamedian uutisia, sekä soveltaa ohjattua koneoppimista eri kehystämisen tapojen tunnistamiseen. Tutkimuskysymyksiin vastaamiseksi kerättiin kattava aineisto eräästä suomalaisesta vastamediasta, ja eroteltiin aineistosta valtamedialinkin sisältävät artikkelit. Tämän jälkeen identifioitiin laadullisesti kolme tapaa jolla vastamedia kehystää valtamedian uutisia: kritisoimalla valtamediaa, kopioimalla sisältöä sekä hyödyntämällä valtamedialähdettä argumentoinnin välineenä.

Tässä tutkielmassa rakennetaan ohjattu koneoppimismalli kolmen edellä luetellun kehystämisen prosessin identifiointiin. Malli rakennettiin luokittelemalla 1000 artikkelin satunnaisotos valtamedialähteen sisältävästä aineistosta kolmeen edellä lueteltuun kehystämisen prosessin kategoriaan. Tämän jälkeen luokitellusta datasta eristettiin erilaisia piirteitä ja rakennettiin näiden pohjalta luokittelija.

Työssä vertailtiin erilaisia satunnaismetsäluokittelijoita sekä tukivektorikoneita, joista eräs satunnaismetsäluokittelija suoriutui luokittelutehtävästi parhaiten. Luokittelijaa ei kuitenkaan voida pitää tarpeeksi tarkkana useimpiin käytännön hyvin korkeaa tarkkuutta vaativiin sovelluksiin. Luokittelijan merkittävimpinä pitämistä piirteistä saadaan kuitenkin uutta tietoa siitä, miten eri sanoja ja tekstin muotoilutyylillisiä keinoja käytetään eri kehystämistavoissa. Esimerkiksi artikkeleissa käytettyjen linkkien määrä sekä alaotsikkojen määrä nousivat luokittelijalle merkittävimpien piirteiden joukkoon.

Tuloksista voidaan päätellä, että laskennallisessa mediatutkimuksessa sanojen lisäksi on hyödyllistä eristää myös artikkeliin liittyvää muotoiludataa. Toinen keskeinen tulos on, että ohjattua koneoppimista voidaan hyödyntää erilaisten median lähteeseen suuntautuvien orientaatioiden tunnistamiseen.

Avainsanat — Nyckelord — Keywords
laskennallinen kehysanalyysi, ohjattu koneoppiminen, vastamedia

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

1 Johdanto

Valeuutiset ovat viime aikoina nousseet merkittäväksi yhteiskunnallisen keskustelun aiheeksi, kansainvälisesti erityisesti vuoden 2016 Yhdysvaltain presidentinvaaleihin liittyen [Allcott ja Gentzkow, 2017]. Vaalien aikana joitakin valeuutisia levitettiin laajemmalle kuin suosituimpia valtamediauutisia, ja on myös arveltu että valeuutiset auttoivat merkittävästi Trumpin vaalivoittoa [Allcott ja Gentzkow, 2017].

Valeuutiselle ei ole yksikäsitteistä määritelmää eikä ole olemassa yhtä keinoa valeuutisen tunnistamiseen. Allcott and Gentzkow määrittelevät valeuutiset "uutisartikkeleiksi, jotka ovat tarkoituksellisesti and todistettavasti valhetta, ja voivat harhaanjohtaa lukijaa" [Allcott ja Gentzkow, 2017]. Viimeaikaisessa tutkimuksessa valeuutinen on määritelty myös misinformaation ja disinformaation käsitteiden kautta erottelemaan kirjoittajan tarkoitusperiä [Wu et al., 2014]. Misinformaatio on väärää tietoa jota on tahattomasti jaettu, kun taas disinformaation on tiedetty olevan valhetta sitä jaettaessa [Wu et al., 2014].

Laskennallinen valeuutistutkimus voidaan jakaa niiden sisällön sekä lukijakunnan tutkimiseen. Esimerkiksi Yhdysvaltalaisesta lukijatutkimuksesta tiedetään, että valeuutissivustoilla vierailee hyvin vähän lukijoita verrattuna valtamediasivustoihin [Nelson ja Taneja, 2018], ja että valeuutissivustoilla vieraillessa vietetään vähemmän aikaa kuin valtamediasivustoilla vieraillessa. Valeuutissivustoille tiedetään myös tulevan enemmän vierailuja Facebookista kuin valtamediasivustoille [Nelson ja Taneja, 2018].

Valeuutisten sisällön laskennallinen analyysi on keskittynyt lähinnä valeuutisten tunnistamiseen, eli uutisten luokittelemiseksi todeksi tai valheeksi. Valeuutisten tunnistamiseen käytettävät algoritmiset lähestymistavat voidaan jakaa karkeasti kahteen ryhmittymään: algoritmit, jotka tunnistavat valeuutisia niiden lingvististen ominaisuuksien perusteella [Pérez-Rosas et al., 2017], ja algoritmit, jotka tunnistavat valeuutisia niiden jakoverkoston muodon avulla [Kumar ja Geethakumari, 2014, Shu et al., 2019]. Lingvististen ominaisuuksien perusteella tunnistamisessa on käytetty esimerkiksi n-grammeja ja psykolingvistisiä ominaisuuksia [Pérez-Rosas et al., 2017], jakoverkoston perusteella tunnistaessa gini-kerrointa [Kumar ja Geethakumari, 2014].

Suomessa valemedioita kutsutaan myös vastamedioiksi, koska monet niiden artikkeleista eivät sisällä väärää tietoa [Noppari ja Hiltunen, 2017]. Tiedetään myös, että Suomessa vastamediauutiset eivät toimi erillisenä osana mediaekosysteemiä, vaan ne kehystävät valtamedian uutisia tukemaan omia agendojaan [Ylä-Anttila, 2018, Noppari ja Hiltunen, 2017]. Tästä seuraa, että erityisesti suomalaisessa ympäristössä vastamedioiden sisältöä tulisi ymmärtää myös muilla mittareilla kuin todeksi tai valheeksi luokittelemalla, esimerkiksi selvittämällä, miten vastamediat kehystävät valtamediauutisia.

Kehystämisellä tarkoitetaan viestinnän tutkimuksessa prosessia, jossa valintojen ja jäsentämisen avulla muokataan mediaesityksen tulkintaa [Seppä-

nen, 2014, s. 97]. Kehystämisessä voidaan käyttää valinnan ja poissulkemisen lisäksi esimerkiksi metaforia ja iskulauseita [Väliverronen, 1996, s. 111]. Kehyksen käsite on alun perin lähtöisin sosiologi Erving Goffmanilta [Goffman, 1974], ja kehysanalyysiin on sovellettu viestintätieteellisessä tutkimuksessa monenlaisia menetelmiä [Väliverronen, 1996, 108-111]. Esimerkiksi journalismissa yksi yleinen kehys on kiistan kehys, jonka sisältävissä uutisissa on havaittavissa vastakkainasetelma toimijoiden välillä [Väliverronen, 1996, 109].

Datatieteessä ei ole yhtä vakiintunutta menetelmää automaattiseen kehysanalyysiin, koska kehysanalyysiä on hankala operationalisoida yksikäsitteisesti laskennalliseen muotoon. Automaattista kehysanalyysiä onkin tehty aiemmin sekä ohjatuilla [Burscher et al., 2014], että ohjaamattomilla [Pashakhin, 2016] koneoppimismenetelmillä. Ohjatuilla koneoppimismenetelmillä on esimerkiksi luokiteltu uutisia joihinkin yleisiin uutisten kehyksiin, esimerkiksi konfliktikehykseen, taloudellisten seurausten kehykseen, ihmiskeskeiseen kehykseen ja moraaliseen kehykseen, käyttäen erilaisten tukivektorikoneiden ja perseptronialgoritmin yhdistelmäluokittelijaa [Burscher et al., 2014]. Ohjaamattomilla koneoppimismenetelmillä taas on yritetty tunnistaa yleisiä kehyksiä esimerkiksi pääkomponenttianalyysillä [Greussing ja Boomgaarden, 2017], klusteroinnilla [Burscher et al., 2016] ja aihemallinnuksella [Pashakhin, 2016].

Tämän tutkielman tutkimuskysymykset ovat:

1. Tutkimuskysymys 1: Miten suomalaiset vastamediat kehystävät valtamedian uutisia?

2. Tutkimuskysymys 2: Millaisella menetelmällä valtamediauutisten kehystämisen tunnistaminen voidaan automatisoida?

Koska laskennalliseen kehysanalyysiin ei ole olemassa tähän tutkimuskysymykseen sopivaa vakiintunutta menetelmää, toiseen tutkimuskysymykseen vastaamiseksi rakennetaan ohjattu koneoppimismalli.

Tämän tutkielma sijoittuu laskennallisen mediatutkimuksen alueelle. Taustaksi valeuutisilmiöstä, toisessa luvussa esitellään valeuutiskäsite sekä suomalaista vastamediakeskustelua, sekä kolmannessa luvussa valeuutisten laskennallista tutkimusta. Luvussa 4 esitellään yhteiskuntatieteellistä kehysanalyysiä ja sen laskennallisia lähestymistapoja taustaksi kehysanalyysin aiemmasta tutkimuksesta. Viidennessä luvussa käsitellään ohjatun koneoppimisen perusteoriaa tekstianalyysiin sovellettuna. Tutkimuskysymyksiin vastaamiseksi rakennettu koneoppimismalli sekä siihen liittyvät tulokset esitellään luvuissa 6 ja 7. Lopuksi kahdeksannessa luvussa käsitellään mallin ja datankeruun rajoitteita sekä tulevaisuuden tutkimuskohteita.

2 Valeuutinen ja vastamedia

Valeuutisen käsitettä käytetään eri merkityksissä, ja sitä on sen vuoksi myös kritisoitu. Suomessa valemedioita kutsutaan myös vastamedioiksi. Tässä luvussa esitellään valeuutisen ja vastamedian käsitteiden merkityksiä ja taustaa.

Valeuutiselle ei ole yksikäsitteistä määritelmää, ja sitä on käytetty hyvin erilaisissa merkityksissä. Ennen Yhdysvaltain vuoden 2016 presidentinvaaleja valeuutisista puhuttiin enimmäkseen tutkittaessa uutisparodiaa ja -satiiria. Esimerkiksi Marchi kirjoitti vuonna 2012 valeuutisilla viittaavansa tv-ohjelmiin, jotka käyttävät satiiria parodioidakseen internet-uutisia [Marchi, 2012]. Samana vuonna esimerkiksi Day & Thompson tutkivat valeuutisia käsitellen Saturday Night Liven parodiaa esittävää "Weekend's update" -ohjelmaa [Day ja Thompson, 2012].

Yhdysvaltain vuoden 2016 presidentinvaalien aikaan ja jälkeen valeuutisen käsite on laajentunut, ja niistä on alettu kirjoittaa myös viitaten sosiaalisessa mediassa kiertäviin erilaisiin virheellisiin artikkeleihin, jotka voivat harhaanjohtaa lukijaa [Allcott ja Gentzkow, 2017]. Presidentinvaalien aikaan kiertäneet valeuutiset muistuttavat oikeita uutisia eikä niitä voida tunnistaa satiiriksi tai parodiaksi samalla tavalla kuin aiemman määritelmän mukaisia valeuutisia. Esimerkiksi Facebookissa eniten huomiota saaneessa valeuutisessa kerrottiin Paavi Francisin tukevan Trumpia presidentiksi [Silverman, 2016].

Vaalien aikaan julkaistiin sekä Trumpia että Clintonia koskevia virheellisiä artikkeleita [Subramanian, 2017], ja viimeisten kolmen kampanjakuukauden aikana suosituimmat valeuutiset saivat enemmän huomiota Facebookissa kuin suosituimmat valtamediauutiset [Silverman, 2016]. Facebookissa suurin osa suosituimpien valeuutisten sisällöstä suosi Trumpia [Silverman, 2016], ja onkin arveltu että Trump ei olisi voittanut ilman valeuutisista saamaansa hyötyä [Allcott ja Gentzkow, 2017].

Valeuutinen -termin käyttöä on viime aikoina kritisoitu ainakin kahdesta syystä. Ensimmäiseksi, termi valeuutinen ei kuvaa ilmiön monimutkaisuutta: valeuutisilmiö kattaa myös sisällön joka on vain osittain valheellista, ja ilmiöön kuuluu olennaisena osana niiden jakajakunta, joka ei välttämättä osaa arvioida tiedon paikkaansapitävyyttä [European commission, 2018]. Toiseksi, valeuutiskäsitettä käytetään myös leimaamaan sellaista sisältöä jonka kanssa ei olla samaa mieltä, vaikka itse sisältö ei olisikaan valheellista [European commission, 2018].

Euroopan Neuvosto on vastannut valeuutiskäsitteen ongelmiin kehittämällä konseptuaalisen kehikon, jolla valeuutisilmiötä, tai raportin mukaan,"sisältösekaannusta", voidaan ymmärtää paremmin erilaisten jaotteluiden avulla [Wardle ja Derakhshan, 2017]. Kehikko erottelee informaation tyypin, informaation levittämisen sekä ilmiössä mukana olevat toimijat kolmeen kategoriaan, joista tyyppijaottelun esittely on tämän tutkielman kannalta

olennaista. Kehikossa informaatio jaotellaan tyypiltään mis-, dis- ja malinformaatioon. Misinformaatiolla tarkoitetaan virheellistä tietoa, jolla ei ole haluttu vahingoittaa, esimerkiksi sosiaalisessa mediassa jaettua väärää tietoa jota ei jakohetkellä ole tiedetty vääräksi. Disinformaatio on tarkoituksella vääräksi luotua tietoa, esimerkiksi vaalikampanjoissa tarkoituksella luotuja ja jaettuja vääriä huhuja. Malinformaatiolla tarkoitetaan oikeaa tietoa jota kuitenkin käytetään vahingoittamaan muita [Wardle ja Derakhshan, 2017]. Misinformaation ja disinformaation käsitteitä käytetään aktiivisesti tutkimuksessa valeuutisten rinnalla ja niiden korvaajana [Wu et al., 2014].

Suomessa valemedioita kutsutaan myös vastamedioiksi ja näinollen valeuutisia vastamediauutisiksi, koska monet valemedioiksi miellettyjen sivustojen uutisista eivät sisällä väärää tietoa [Noppari ja Hiltunen, 2017]. Tämän sekä valeuutiskäsitteen ongelmallisuuden vuoksi tässä tutkielmassa käytetäänkin suomalaisista valeuutisista puhuttaessa termiä vastamediauutinen.

Suomalaisesta tutkimuksesta tiedetään myös, että vastamediauutiset eivät toimi erillisenä osana mediaekosysteemiä, vaan ne kehystävät valtamedian uutisia tukemaan omia agendojaan [Ylä-Anttila, 2018, Noppari ja Hiltunen, 2017]. Usein vastamediauutisessa onkin valtamedian jutun tarina esimerkiksi uudella otsikolla tai johdannolla [Noppari ja Hiltunen, 2017]. Valtamediauutisten uudelleenkehystämiseen liittyy omat ongelmansa vaikka itse sisältö olisikin totta: lopputuloksena syntyneestä vastamediauutisesta voi olla hankalaa erottaa mikä osa on alkuperäisestä valtamediauutisesta, ja mikä vastamedian lisäämää tekstiä [Noppari ja Hiltunen, 2017]. Uutisten suoraan kopioimiseen liittyy myös tekijänoikeudellisia ongelmia.

3 Laskennallinen valeuutistutkimus

Laskennallinen valeuutistutkimus voidaan jakaa niiden sisällön, lukija- ja jakajakunnan sekä valeuutisten leviämisen tutkimukseen. Tässä luvussa esitellään valeuutisten laskennallista tutkimusta, johon kuuluu valeuutisten sekä väärän tiedon tunnistamisen ja erottelun menetelmiä. Menetelmät hyödyntävät koneoppimista, ja tämä luku onkin jaettu ohjattuihin ja ohjaamattomiin menetelmiin. Ohjattujen menetelmien luvussa käsitellään lähinnä valeuutisten tunnistamista, ohjaamattomien menetelmien luvussa väärän ja oikean tiedon erottelua sekä misinformaation jaottelua eri kategorioihin.

3.1 Ohjatut valeuutisten tunnistusalgoritmit

Valeuutisten sisällön laskennallisen tutkimuksen yleinen tutkimuskysymys on, miten tunnistaa valeuutinen muista uutisista. Tähän kehitetyt ohjatut menetelmät hyödyntävät koneoppimista ja luonnollisen kielen käsittelyä [Kumar ja Geethakumari, 2014]. Ohjatut menetelmät perustuvat uutisten sisällön lingvististen ominaisuuksien analyysiin [Qazvinian et al., 2011] sekä uutis-

ten Twitterissä tapahtuneiden jakojen jakajaverkoston analyysiin [Kumar ja Geethakumari, 2014, Shu et al., 2019].

Esimerkiksi Qazvinian, Rosengren, Radev ja Mei ovat yrittäneet tunnistaa ristiriitaista tietoa, ”huhuja”, twiiteistä kehittämällä ohjatun menetelmän, käyttäen piirteinä twiittien lingvistisiä ominaisuuksia, niiden jakoverkoston ominaisuuksia sekä twiittien hashtageja ja linkkejä [Qazvinian et al., 2011]. Luokittelija rakennettiin luokittelemalla twiittejä huhuja sisältäväksi ja sisältämättömiksi twiiteiksi, ja sen jälkeen eristämällä piirteitä twiiteistä ja niiden jakoverkostosta. Jokaiselle piirteelle muodostettiin Bayes-luokittelija ja näille log-lineaarinen malli.

Twiittien sisällöstä Qazvinian, Rosengren, Radev ja Mei käyttivät koneoppimismallin piirteinä tokenisoituja sanoja ja sanaluokkia sekä unigrammeina että bigrammeina. Twiittien jakoverkostoon perustuen he rakensivat ensin kaksi todennäköisyysjakaumaa: positiivisen ja negatiivisen jakauman, joista positiivinen kuvasi käyttäjäkohtaisesti twiitin todennäköisyyttä olla huhutwiitti ja negatiivinen todennäköisyyttä olla ei-huhutwiitti. Tämän jälkeen he käyttivät jakaumia apuna kahden piirteen määrittelyssä. Hashtageja ja linkkejä he käyttivät rakentaen todennäköisyysjakaumat hashtageista sekä linkeistä erikseen käyttäen niitä piirteinä. Mallin tarkkuus ylsi twiittien luokittelussa yli 90 prosenttiin [Qazvinian et al., 2011].

Kumar ja Geethakumari ovat kehittäneet käsitteellisen kehikon misinformaation tunnistamiseen, sekä laskennallisen misinformaatiotunnistusmenetelmän siihen perustuen [Kumar ja Geethakumari, 2014]. Kehikko perustuu neljään kognitiivisesta psykologiasta lainattuun tapaan jolla ihminen arvioi tiedon paikkaansapitävyyttä: tiedon johdonmukaisuuteen eli yhtenäisyyteen lukijan aiemman tietämyksen kanssa, tiedon sisäiseen yhtenäisyyteen ilman ristiriitoja, lähteen luotettavuuteen sekä yleiseen hyväksyttävyyteen eli siihen, pitävätkö muut ihmiset tietoa paikkaansapitävänä [Kumar ja Geethakumari, 2014].

Kumar ja Geethakumari operationalisoivat kehikkonsa neljä informaation arviointitapaa laskennallisiksi lähestymistavoiksi, ja rakensivat kahden näistä pohjalta menetelmän misinformaation tunnistamiseen [Kumar ja Geethakumari, 2014]. He ottivat tutkimuksensa kohteeksi vain uudelleentwiittaukset, yhtenä syynä uudelleentwiittausten suurempi mahdollisuus levitä laajalle ja näinollen niiden suurempi merkittävyys misinformaation tutkimuksessa.

Kumar ja Geethakumari operationalisoivat tiedon lähteen luotettavuuden gini-kertoimen avulla. Gini-kertoimella mitataan jakauman epätasaisuutta, ja tässä tapauksessa sillä arvioitiin yhden käyttäjän twiittien uudelleentwiittauksista muodostunutta jakaumaa. Pieni, lähellä nollaa oleva gini-kertoimen arvo merkitsi sitä että uudelleentwiittaukset tulivat pieneltä joukolta käyttäjiä verrattuna twiittien määrään, suuri, lähellä yhtä oleva gini-kertoimen arvo taas sitä että uudelleentwiittaukset hajautuivat monelle käyttäjille. Kumar ja Geethakumari tekivät oletuksen siitä, että tieto on todennäköisemmin misinformaatiota, jos se on peräisin käyttäjältä jonka twiittejä on

uudelleentwiitannut vain pieni joukko.

Tiedon yleisen hyväksyttävyyden Kumar ja Geethakumari operationalisoivat PageRank-algoritmin avulla [Kumar ja Geethakumari, 2014]. PageRank on algoritmi, joka on alun perin kehitetty verkkosivustojen pisteytykseen esimerkiksi hakukoneita varten. Se mittaa sivustolle vieviä linkkejä huomioiden myös linkit sisältävien sivustojen pisteytyksen [Page et al., 1999]. Kumar ja Geethakumari sovelsivat PageRankia twiittikohtaisesti niin, että twiitin pisteytys riippui sen alkuperäisen twiittaajan muiden twiittien suosiosta, sekä twiitin uudelleentwiitanneista käyttäjistä [Kumar ja Geethakumari, 2014]

Kumar ja Geethakumari rakensivat menetelmänsä siten, että se arvioi ensin twiitin lähteen luotettavuutta gini-kertoimen avulla, ja mikäli lähteen luotettavuus on suuri, sen jälkeen yleistä hyväksyttävyyttä PageRank -algoritmin avulla. Mikäli lähteen luotettavuus on pieni, twiitin luokiteltiin olevan mahdollista disinformaatiota vaikka sen yleinen hyväksyttävyys olisikin suuri [Kumar ja Geethakumari, 2014]. Kumar ja Geethakumari eivät arvioineet koko kaksivaiheisen menetelmänsä tarkkuutta, mutta lähteen luotettavuutta arvioinut gini-kerrointa käyttänyt algoritmi saavutti 90 prosentin tarkkuuden misinformaation tunnistamisessa [Kumar ja Geethakumari, 2014].

Vaikka ylläesiteltyjen virheellisen tiedon tunnistusmenetelmien tarkkuudet ovat hyvällä tasolla, menetelmät eivät kuitenkaan ole ongelmattomia. Lingvistisiin ominaisuuksiin perustuvissa luokittelijoissa tulee huomioida, että kirjoitustavat muuttuvat ajan myötä, joten luokittelijoiden saavutettu tarkkuus ei välttämättä päde uudella aineistolla ajan myötä [Edell, 2018]. Molemmat ylläesitellyt menetelmät myös sisältävät ihmisen arviointia siitä mitä on virheellinen tieto. Esimerkiksi Kumarin ja Geethakumarin lähteen luotettavuudeksi operationalisoima gini-kerroin vaatii käytössä jonkunlaisen rajan, mikä alitettaessa tehdään päätös tiedon virheellisyydestä.

Mikäli gini-kertoimen misinformaatioraja opitaan misinformaatioksi luokitelluista opetusdatatwiiteistä, ihmisen tekemän arvioinnin ongelma siirtyy opetusdatan luokitteluvaiheeseen. Kuten edellisessä luvussa 2 todettiin, valeuutisen tai tässä tapauksessa misinformaation käsitettä ei voida yksikäsitteisesti määritellä, joten luokittelut ovat väistämättä jossain määrin subjektiivisia. Luokittelun subjektiivisuuden vähentämiseen on kuitenkin keinoja: esimerkiksi Qazvinian, Rosengren, Radev ja Mei luokittelivat 500 twiittiä kahden ihmisen voimin, ja määrittivät luokituksille Kappa-arvon, joka mittaa havaintojen yhteneväisyyttä [Qazvinian et al., 2011], arvioidakseen luokitusten subjektiivisuutta. Kappa-arvon laskentaperiaate esitellään tarkemmin tämän tutkielman luvussa 5.2.1.

3.2 Ohjaamattomat valeuutisten ja väärän tiedon tunnistus- ja luokittelumenetelmät

Virheellisen ja oikean tiedon erotteluun on myös kehitetty menetelmiä, jotka eivät vaadi aluksi ihmisen tekemää opetusdatan luokittelua. Menetelmiä joiden tarkoitus on erotella yleisesti oikeaa ja väärää tietoa, ei välttämättä valeuutisia tai misinformaatiota, kutsutaan usein *truth discovery* -menetelmiksi [Wang et al., 2012, Yin et al., 2008]. Näissä menetelmissä on hyödynnetty ohjaamattomia menetelmiä [Wang et al., 2012, Yin et al., 2008].

Esimerkiksi Yin, Han ja Yu [Yin et al., 2008] ovat kehittäneet menetelmän, joka yrittää etsiä ennalta määriteltyyn kysymykseen oikean vastauksen hyödyntäen verkkosivujen ja niillä esiintyvien faktojen suhdetta. Menetelmän perusoletuksena on, luotettavilla verkkosivustoilla esiintyy paljon faktatietoja, ja tiedot ovat faktatietoja mikäli ne esiintyvät monilla luotettavilla verkkosivustoilla. Tieto ei todennäköisesti ole faktaa, jos se on ristiriidassa monen luotettavalla sivustolla esiintyvän tiedon kanssa. Yin, Han ja Yu kehittivät iteratiivisen menetelmän tiedon totuudellisuuden arviointiin [Yin et al., 2008]. Menetelmä saavutti korkean tarkkuuden ennalta määritellyssä ongelmassa kirjojen kirjoittajien päättelyssä internetistä. Ongelmana tiedon laajalle levinneisyyteen perustuvissa menetelmissä kuitenkin on, että joskus myös väärä tieto saattaa levitä laajemmalle kuin oikea tieto samasta aiheesta [Yin ja Tan, 2011].

Hosseinimotlagh ja Papalexakis kehittivät ohjaamattoman koneoppimisen menetelmän ennalta määriteltyjen valeuutistyyppien, esimerkiksi satiirin ja salaliittoteorioiden tunnistamiseen [Hosseinimotlagh ja Papalexakis, 2018]. Heidän aineistossaan artikkelit oli koodattu eri tyyppeihin, mutta tyyppejä ei käytetty ohjaamattoman algoritmin syötteenä. He vertailivat tutkimuksessaan eri ohjaamattomia koneoppimismenetelmiä, joista osa analysoi artikkeleita niiden sanojen esiintymismäärien perusteella, ja osa sekä sanaesiintymisten että sanojen vierekkäisyyksien perusteella (CP/PARAFAC -tensorihajotelma). Paras tarkkuus eri valeuutistyyppien tunnistamisessa saatiin CP/PARAFAC -tensorihajotelmalla, joka on eräänlainen matriisin tekijöihinjakomenetelmä [Hosseinimotlagh ja Papalexakis, 2018]. Myös tässä menetelmässä, kuten viime aliluvussa esitellyissä lingvistisia ominaisuuksia hyödyntävissä ohjatuissa koneoppimismenetelmässä, ongelmana on, että koneoppimismalli on rakennettu tietyn aineiston pohjalta, eivätkä tutkimuksen tekijät ole validoineet malliaan uudella tai erilaisella aineistolla.

4 Kehysanalyysi mediatutkimuksessa

Kehysanalyysi on yhteiskuntatieteellinen menetelmä, jota käyttäväksi on nimetty monenlaisia tutkimuksia. Kehysanalyysi sai alkunsa ihmisten vuorovaikutuksen tutkimuksesta, ja myöhemmin sitä on käytetty myös journalismin tutkimuksessa. Tässä luvussa esitellään aluksi yhteiskuntatieteellisiä

kehykselle annettuja määritelmiä. Sen jälkeen esitellään manuaalista yhteiskuntatieteellistä kehysanalyysiä käyttäneitä tutkimuksia, ja tämän jälkeen laskennallista kehysanalyysiä.

4.1 Kehyksen monet määritelmät

Sosiologi Erving Goffman [Goffman, 1974] kehitti kehyksen (frame) käsitteen tutkiessaan ihmisten välistä vuorovaikutusta 1970-luvulla. Hän määrittelee kehyksen tulkintakehikoksi jonka avulla asioita käsitetään: esimerkiksi uuteen tilanteeseen tullessaan ihminen saattaa kysyä: ”mitä täällä on meneillään?”.Vastaus kysymykseen riippuu kuitenkin siitä, miten vastaaja kokee tilanteen. Goffman kutsuukin kehykseksi kokijan tilanteesta havaitsemia peruselementtejä [Goffman, 1974].

Kehyksen käsite tarjoaa myös väylän tutkia viestinnän, esimerkiksi journalismin vaikutusta ihmisen tietoisuuteen [Entman, 1993, s. 51]. Mediatutkimuksessa uutisten kehysanalyysiä on käytetty ainakin neljään eri tarkoitukseen: 1) tunnistettu kehyksiä aineistosta, 2) tutkittu kehysten tuottamisen olosuhteita, 3) tutkittu miten ihmisen aiemmat käsitykset vaikuttavat kehyksen tulkintaan, sekä 4) tutkittu miten uutiskehykset muokkaavat sosiaalisia todellisuutta, esimerkiksi julkisia mielipiteitä [D’Angelo, 2002, s. 873]. Mediatutkimuksessa kehykselle ja kehysanalyysille on annettu erilaisia tarkempia sekä löyhempiä määritelmiä, joista tässä luvussa voidaan käsitellä vain pieni osa. Yhteiskuntatieteelliset tekstitutkimuksen menetelmät kehittyvät aineiston ja tutkimuskysymysten mukaan [Horsti, 2005, s. 48], ja esimerkiksi jotkut suomalaiset kehysanalyysiä soveltaneet tutkijat ovatkin kertoneet noudattavansa vain löyhästi muiden kehittämiä kehysanalyysin määritelmiä [Horsti, 2005, Väliverronen, 1996].

Kehystämistä journalismissa tutkinut Robert Entman määrittelee kehystämisen olevan ”joidenkin näkökantojen valitsemista ja tekemistä merkittävämmäksi tekstillä kommunikoiden” [Entman, 1993, s. 52]. Hänen mukaansa kehystäminen koostuu siis valikoinnista ja merkitysten antamisesta, ja kehykset ilmenevät hänen mukaansa esimerkiksi avainsanojen, lähteiden tai kuvien kautta. Esimerkiksi kehyksestä hän antaa Yhdysvalloissa ulkopoliittista uutisointia menneinä vuosikymmeninä hallinneen kylmän sodan kehyksen, jossa ulkomailla tapahtuneita sisällisssotia kehystettiin kommunistien syyksi ja Yhdysvaltoja kiiteltiin sodan toisen osapuolen auttajana [Entman, 1993, s. 52]. Entman nimeää myös viestinnän tekemisen ja tutkimuksen alueita jotka voisivat hyötyä kehysanalyysistä: esimerkiksi journalistit voisivat kehysanalyysiä tuntiessaan haastaa vallitsevia kehyksiä, ja sisällönanalyysissä voitaisiin välttää arvottavia termejä puhumalla kehyksistä [Entman, 1993, s. 56-57].

Kehysanalyysin teoretisointiin liittyy perustavanlaatuinen ongelma siitä, missä määrin kehykseen vaikuttaa tekstin lukija tai tilanteen kokija, ja missä määrin kehys on ympäristön tai tekstin kirjoittajan määrittelemää.

D'Angelo onkin jaotellut kehysanalyysin kolmeen eri paradigmaan osittain tämän ominaisuuden perusteella [D'Angelo, 2002]. Ensimmäinen paradigma on kognitiivinen paradigma, jonka mukaan tehdyissä tutkimuksissa ollaan kiinnostuneita ihmisen semanttisesta ymmärryksestä ja esimerkiksi siitä, miten ihmisen aiempi tietämys vaikuttaa kehysten ymmärtämiseen [D'Angelo, 2002, s.875]. Toisen paradigman, kriittisen paradigman, mukaiset tutkijat näkevät kehysten muodostuvan journalistisen koostamisen tuloksena siten, että kehykset mukailevat poliittisen ja taloudellisen eliitin arvoja [D'Angelo, 2002, s.876]. Kolmas paradigma, konstruktivistinen paradigma, näkee journalistit "tiedonkäsittelijöinä", jotka muotoilevat tulkinnallisia kehyksiä [D'Angelo, 2002, s.877].

Konstruktionistien ja kriittisen paradigman toteuttajien erona on esimerkiksi se, että konstruktionistit näkevät journalistisen lähteiden valitsemisen tapahtuvan sen mukaan mitä hyviä lähteitä on tarjolla, kun taas kriittisen paradigman toteuttajat näkevät lähteiden valinnan tapahtuvan yhteiskunnan vallankäyttäjien ehdoilla [D'Angelo, 2002, s.877]. D'Angelo nimeää Goffmanin yhdeksi konstruktionistista paradigmaa toteuttavaksi tutkijaksi, koska Goffman on käsitellyt tutkimuksessaan paljon tiettyjä kehyksiä ja sitä miten niitä voi tunnistaa, ei niinkään institutionaalisia prosesseja jotka tuottavat kehyksiä [D'Angelo, 2002, s.878].

Gamson ja Lasch nimittävät kehyksiä tulkintapaketeiksi, ja he ovat artikkelissaan *The Political Culture of Social Welfare Policy* kehittäneet teoreettisia välineitä analysoida tulkintapaketteja [Gamson ja Lasch, 1983]. Tulkintapaketti koostuu heidän mukaansa kehyksen määrittelevästä tilanteesta sekä itse kehyksestä. Tulkintapaketilla on joukko elementtejä jotka määrittelevät sen kehyksen ja sen määrittelevän tilanteen, ja näitä elementtejä kutsutaan *allekirjoitukseksi* (signature). Allekirjoitukseen kuuluu kehystämisen (framing) ja perustelun (reasoning) välineitä, jotka muodostavat kehyksen olemuksen. Kehystämisen välineitä ovat esimerkit, metaforat, kuvaukset, visuaaliset kuvat ja tunnetut lauseet. Perustelun välineitä taas ovat "syyt, seuraukset ja periaatteisiin vetoaminen" [Gamson ja Lasch, 1983]. Journalistiset tulkintapaketit muodostuvat vuorovaikutuksessa journalistien ja lähteiden kesken.

Gamson ja Lasch esittävät yhdeksi tavaksi analysoida tulkintapaketteja muodostaa *allekirjoitusmatriisin*, jossa rivit kuvastavat eri tulkintapaketteja, ja kolumnit allekirjoituksen elementtejä. Kolumneihin kirjoitetaan eri tulkintapaketeille aineistosta löytyviä ominaisia elementtejä, esimerkiksi metaforia metaforien kolumniin. Gamson ja Lasch sovelsivat tulkintapakettien teoreettista kehikkoaan hyvinvointivaltiopolitiikasta kertovien mediakehysten tutkimiseen [Gamson ja Lasch, 1983]. Tähän tutkimukseen palataan seuraavassa aliluvussa.

4.2 Manuaaliset kehysanalyysimenetelmät

Kehysanalyysiä on tehty monenlaisissa tutkimuksissa [Horsti, 2005, s. 50]. Esimerkiksi Esa Väliverronen [Väliverronen, 1996] on soveltanut kehysanalyysiä metsätuhouutisoinnin tutkimiseen. Aineistonaan hänellä oli 103 suomalaista uutisartikkelia 1980- ja 1990 -lukujen vaihteesta [Väliverronen, 1996, s. 89], jolloin metsätuhot olivat tärkeä poliittinen teema [Väliverronen, 1996, s. 71]. Väliverronen teki aineistolleen sekä sisällönanalyysiä esimerkiksi uutisissa esiintyvistä toimijoista, että varsinaista kehysanalyysiä. Kehysanalyysissään hän otti vaikutteita Gamsonin ja Laschin edellisessä aliluvussa esitellystä teoreettisesta tulkintapakettikehikosta, esimerkiksi luonnostelemalla allekirjoitusmatriisin löytämilleen kehyksille [Väliverronen, 1996, s. 111-113].

Väliverronen löysi aineistostaan kolme kehystä: sairauden, kiistan ja hallinnan kehykset [Väliverronen, 1996, s. 112]. Sairauden kehyksen sisältävissä uutisissa metsätuhot esitettiin ”metsän sairastumisen” eli ihmisen aiheuttaman saastumisen seurauksena. Kiistan kehyksessä uutisoitiin metsätuhoihin liittyvien toimijoiden, esimerkiksi tutkijoiden välisistä erimielisyyksistä. Hallinnan kehyksessä metsätuhot esitettiin hallittavissa olevana uhkana, johon tiede tulisi löytämään ratkaisuja [Väliverronen, 1996, s. 112-114]. Sairauden ja kiistan kehykset Väliverronen tunnisti helposti omasta sanastostaan ja visuaalisista kuvista [Väliverronen, 1996, s. 120]: esimerkiksi Lapin Kansa uutisoi että ”Lapista ei löydy yhtään tervettä neulasta” [Väliverronen, 1996, s. 117], ja kiistan kehyksessä Demari uutisoi että ”Metsäntutkijoiden saastesoppa kiehuu” [Väliverronen, 1996, s. 118].

Hallinnan kehykseen ei liittynyt yhtä tunnistettavaa kielenkäyttöä, vaan siihen kuuluvat uutiset olivat osa pidempiaikaista muutosta metsätuhojen näkemiseen ratkaistavissa olevana tutkimuksellisena ja hallinnollisena ongelmana [Väliverronen, 1996, s. 120-121]. Hallinnan kehys on tunnistettu myös muissa yhteiskunnallisten ilmiöiden uutisoinnin kehysanalyysiä tehneissä tutkimuksissa: esimerkiksi Karina Horstin turvapaikanhakijauutisointia koskeneessa tutkimuksessa yksi kehyksistä oli hallinnan kehys [Horsti, 2005, s. 141]. Horstin mukaan hänen tutkimuksessaan hallinnan kehys näkyykin metaforien ja iskulauseiden sijaan enemmän perustelun keinoissa, esimerkiksi viranomaiskäsitteiden esiintymisessä tekstissä [Horsti, 2005, s. 144].

Gamson ja Lasch [Gamson ja Lasch, 1983] tutkivat hyvinvointivaltiopolitiikkaa käyttäen tulkintapakettien teoriaansa apuna kehysten määrittelyssä. He tunnistivat esimerkiksi hyvinvoinnin, vapaamatkustajien sekä työtä tekevien köyhien kehykset. Aineistona josta he tunnistivat kehykset, oli esimerkiksi puheita, todistuksia, tiedotteita ja pamfletteja, sekä näyte lehtiartikkeleista. Erona Väliverrosen edellä kuvattuun tutkimukseen, jossa Väliverronen luonnosteli tulkintakehikkonsa suoraan media-aineistostaan, Gamson ja Lasch tunnistivat kehyksensä pääosin muusta aineistosta [Väliverronen, 1996, s. 111]. Heidän mukaansa allekirjoitusmatriisin luonnosteleminen on vasta en-

simmäinen vaihe: seuraava vaihe olisi analysoida järjestelmällisesti, miten tunnistetut kehykset ilmenevät mediateksteissä [Gamson ja Lasch, 1983]. Väliverroselle kehykset ovat siis tutkimustulos, kun Gamsonille ja Laschille ne ovat välivaihe [Väliverronen, 1996, s. 111].

David, Atun, Fille, ja Monterola vertailivat kahta menetelmää Filippiinien väestöpolitiikan mediakehysten selvittämiseen siten, että ensimmäisessä menetelmässä kehykset muodostettiin aineiston pohjalta, ja toisessa aiheen asiantuntijat muodostivat kehykset tietämyksensä ja muiden tekstien kuin aineiston pohjalta [David et al., 2011]. Aineiston pohjalta muodostettaessa kehykset tunnistettiin *Matthesin ja Kohringin menetelmällä*, joka perustuu *kehyselementtien* esiintyvyyksiin teksteissä. Kehyselementtejä olivat tutkimuksessa esimerkiksi aihe, toimija, hyöty ja riski. Jokainen kehyselementti sisälsi binäärisiä muuttujia, esimerkiksi riskielementti ekonomisten riskien ja resurssien loppumisen muuttujat. Jokaiselle binääriselle muuttujalle koodattiin arvo artikkelin sisällön mukaan, ja tämän jälkeen käytettiin klusterointia artikkeleiden luokittelemiseen ryhmiin kehyselementtien muuttujien arvojen perusteella. Klusteroinnista muodostuneet artikkeliryhmät tulkittiin saman kehyksen sisältäviksi, ja ryhmien artikkeleiden perusteella kehyksille annettiin nimet. Toisessa menetelmässä taas jokainen aineiston artikkeli luokiteltiin yhteen valmiin kehyslistan kehykseen [David et al., 2011].

Klusteroinnista syntyi kolme kehystä, kun taas valmiissa kehyslistassa oli viisi kehystä. Menetelmien tulokset olivat samankaltaisia: molempien menetelmien tuloksista löytyi esimerkiksi väestönkasvun kehys. Aineiston pohjalta määrällisen analyysin sekä valmiiden kehysten pohjalta luokittelussa on omat hyvät ja huonot puolensa: Matthesin ja Kohringin menetelmällä tehty kehysanalyysi mahdollistaa kehysten limittymisen analyysin ajan kuluessa kehyselementtien arvoja tarkastelemalla, kun taas valmiit kehykset säästävät aikaa mikäli kehykset selkeästi eroavat toisistaan [David et al., 2011]. Matthesin ja Kohringin menetelmä voi myös soveltua paremmin tilanteeseen, jossa analysoitava ilmiö on niin uusi, että siitä ei ole helppoa tunnistaa kehyksiä ennen aineiston käsittelyä.

4.3 Laskennalliset kehysanalyysimenetelmät

Manuaalisessa, käsin tehdyssä kehysanalyysissä, sekä laskennallisissa kehysanalyysimenetelmissä on omat hyvät ja huonot puolensa. Ihmisen tekemä aineiston huolellinen lukeminen voi paljastaa kehyksistä sellaisia asioita, joita laskennalliset menetelmät eivät huomaisi, esimerkiksi siksi koska laskennalliset kehysanalyysimenetelmät perustuvat usein sanojen esiintyvyysmääriin [David et al., 2011]. Sanoilla voi myös olla tekstissä eri merkityksiä, jota laskennalliset menetelmät eivät välttämättä huomaisi.

Manuaalisia kehysanalyysimenetelmiä käytettäessä on kuitenkin usein melko vaikeaa kuvailla tarkasti, miten kehykset ovat syntyneet [Matthes ja Kohring, 2008], ja niiden kuvaillaan usein vain ”ilmenneen” aineistosta.

Kuten luvussa 4.1 todettiin, jotkut esimerkiksi suomalaiset kehysanalyysitutkijat eivät ole lainkaan sitoutuneet tiettyyn analyysimenetelmään, joten tutkimuksen objektiivisuus voi kärsiä. Manuaaliset kehysanalyysimenetelmät eivät myöskään sovi kovin suurille aineistoille niiden aineistokohtaisen työmäärän vuoksi [David et al., 2011].

Kehysanalyysiin on esimerkiksi objektiivisuuden lisäämiseksi kehitetty paljon erilaisia laskennallisia menetelmiä. Tässä aliluvussa esitellään joitakin laskennalliseen kehysanalyysiin kehitettyjä menetelmiä sekä niistä käytyä keskustelua. Osio on jaettu alilukuihin siten, että aluksi ensimmäisessä aliluvussa käsitellään laskennallisten kehysanalyysimenetelmien teoriaa tietojenkäsittelytieteen näkökulmasta. Sen jälkeen esitellään tutkimuksia, joissa menetelmiä on sovellettu kehysanalyysiin. Viimeisessä aliluvussa käsitellään laskennallisista kehysanalyysimenetelmistä käytyä kriittistä keskustelua.

4.3.1 Menetelmien teoria

Laskennalliseen kehysanalyysiin sovelletut menetelmät perustuvat tietojenkäsittelytieteen ja tilastotieteen keskeisiin konsepteihin. Jotkut laskennallisen kehysanalyysin tutkimukset hyödyntävät suoraan valmiita malleja, kuten aihemallinnusta. Toisissa tutkimuksissa taas tutkijat ovat kehittäneet omia laskennallisen kehysanalyysin menetelmiä yksinkertaisempiin käsitteisiin, kuten korrelaatioon perustuen.

Tässä aliluvussa esitellään kolme laskennalliseen kehysanalyysiin sovellettua menetelmää: Latent Dirichlet Allocation -aihemalli [Blei et al., 2003, Pashakhin, 2016], semanttinen kartta [Hellsten et al., 2010] sekä kehyskartoitus (frame mapping) [Miller, 1997]. Näistä Latent Dirichlet Allocation on tunnettu ja paljon useisiin eri tarkoituksiin käytetty menetelmä. Semanttinen kartta ja kehyskartoitus taas ovat joko osittain tai kokonaan niitä soveltaneiden tutkjoiden kehittämiä menetelmiä.

Miller [Miller, 1997] kehitti vuonna 1997 kehysten tunnistamiseen laskennallisen menetelmän, jota hän kutsui nimellä ”frame mapping”. Menetelmä sai innoitusta Entmanin luvussa 4.1 esitellystä kehystämisen määritelmästä, jonka mukaan kehykset ilmenevät esimerkiksi avainsanojen kautta [Miller, 1997, s. 368]. Menetelmän ideana on etsiä sanoja, jotka esiintyvät usein toistensa kanssa joissakin teksteissä (mutta eivät kaikissa teksteissä), ja muodostaa niiden avulla kehyksiä, kuitenkin niin että sanojen tarkalla tekstikohtaisella järjestyksellä ei ole väliä. Kehys operationalisoitiin siis avainsanojen avulla.

Millerin menetelmää voidaan pitää eräänlaisena aihemallinnuksen ja semanttisen kartan esiasteena, ja se on hyvin samankaltainen menetelmä semanttisen kartan kanssa. Menetelmän perustuu seuraaviin vaiheisiin [Miller, 1997]:

1. Muodostetaan jokaiselle dokumentille sanaesiintyvyyslista laskevassa järjestyksessä

2. Valitaan tietty määrä eniten esiintyviä sanoja myöhempään analyysiin

3. Mikäli sanojen esiintyvyyttä yhdessä halutaan analysoida jonkin muun attribuutin kuin eri dokumenttien suhteen, koodataan ja indeksoidaan aineisto halutulla tavalla

4. Muodostetaan kosinietäisyysmatriisi sanoille. Kosini määritellään seuraavasti sanoille x ja y:

$$\text{kosini}(x, y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}, \tag{1}$$

   missä x_i on sanan x frekvenssi ja y_i sanan y frekvenssi i :nnessä dokumentissa.

5. Generoidaan ominaisvektorit jokaiselle sanalle kuvaamaan sanojen sijaintia koordinaatistossa

6. Mikäli sanoja on liian paljon visualisoitavaksi, klusteroidaan sanoja esimerkiksi hierarkisella klusteroinnilla

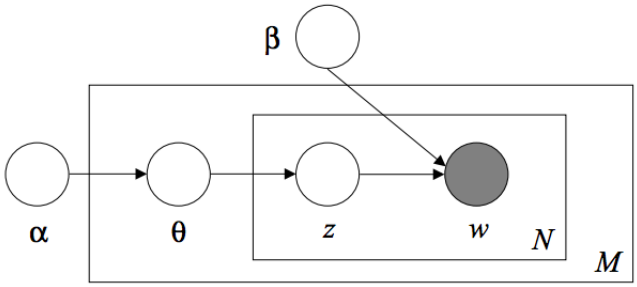
7. Visualisoidaan lopputuloksena olevat sanat koordinaatistossa

Miller otti menetelmässään huomioon myös synonyymit sekä sanat, joilla voi olla useita eri merkityksiä. Hän ehdottaa yleiseksi ratkaisuksi monimerkityksisten sanojen poistamista tai niiden jokaisen esiintymän koodaamista kontekstin mukaan [Miller, 1997].

Latent Dirichlet Allocation (LDA) on diskreettiä dataa varten kehitetty, hierarkinen ja generatiivinen Bayes-malli, jossa on kolme tasoa [Blei et al., 2003]. LDA:ta on kehysanalyysissä käytetty esimerkiksi Ukrainan vuosien 2013-2014 kriisin mediakehysten analyysiin [Pashakhin, 2016]. LDA:ssa datan perusyksikköjä ovat sanat ja dokumentit. Sanat voidaan formalisoida indeksoiduksi sanastoksi $1, ..., V$, dokumentit taas sanojen joukoiksi $(w_1, w_2, ...w_N)$. Myös aiheet ovat LDA:ssa sanojen joukkoja, merkitään yksittäistä aihetta symbolilla z_n.

Kaksi LDA:ssa olennaista jakaumaa, joita kuvaavia parametreja mallissa estimoidaan, ovat aiheiden jakauma dokumenteissa sekä sanojen jakauma aiheissa. LDA:ssa mallin oletuksena on, että dokumentti luodaan siten, että jokaisen sanan kohdalla valitaan ensin aihe. Aiheilla on todennäköisyydet tulla valituksi, ja mallintamisprosessissa niiden todennäköisyysjakauma päätellään datasta. Aiheen valinnan jälkeen aiheesta valitaan sana aihekohtaisesta sanojen todennäköisyysjakaumasta. Formaalimmin ilmaistuna, LDA olettaa että jokainen dokumentti on luotu seuraavanlaisella generatiivisella prosessilla [Blei et al., 2003]:

1. Valitaan sanojen määrä $N \sim Poisson(E)$. Sanojen määrän jakauma on mallin oletus, ja jakauma voidaan valita ilmiötä parhaiten kuvaavaksi.



Kuva 1: Latent Dirichlet Allocation -mallin parametrit sanojen, dokumentin ja korpuksen tasolla [Blei et al., 2003]. Nuolet merkitsevät parametrien ehdollista riippuvuutta.

2. Valitaan $\theta \sim Dir(\alpha)$
3. Jokaiselle sanalle w_n N sanasta:
   - 3.1. Valitaan aihe $z_n \sim Multinomi(\theta)$
   - 3.2. Valitaan sana w_n jakaumasta $p(w_n|z_n, \beta)$, joka kuvaa multinomiaalista sanan todennäköisyyttä aiheelle z_n

Mallille tulee antaa ulkoisena parametrina aiheiden määrä k, joka määrittää Dirichlet-jakauman ulottuvuuden. Lisäksi, sanojen todennäköisyydet muodostetaan $k * V$ matriisiksi β. K-ulotteinen Dirichlet-satunnaismuuttuja määritellään seuraavalla tiheysfunktiolla:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\sum_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} .... \theta_k^{\alpha_k - 1} \tag{2}$$

missä $\Gamma(x)$ on Gamma-funktio [1] ja α on vektori jonka pituus on k.

Näiden lähtökohtien perusteella aiheiden sekoitusta kuvaavan parametrin θ, aiheiden z sekä sanojen w yhteisjakauma määritellään seuraavasti:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3}$$

Yhteisjakaumasta johdetaan dokumentin sekä korpuksen marginaalijakaumat. Kuvasta 1 nähdään, että mallissa on siis estimoitavia parametreja kolmella tasolla: sanojen valinnan (sanakohtainen aihe sekä itse sana aiheen jakaumasta), dokumenttien (dokumenttikohtainen θ) sekä koko korpuksen tasolla (α ja β). Parametrien päättely vaatii approksimoivia päättelyalgoritmeja, kuten esimerkiksi Monte Carlo -menetelmiä [Blei et al., 2003].

[1] Gamma-funktio esitelty tarkemmin [Olver et al., 2019]

Semanttinen kartta on Hellstenin, Dawsonin ja Leydesdorfin kehittämä menetelmä laskennalliseen kehysanalyysiin [Hellsten et al., 2010]. Menetelmän lähtökohtana on sen kehittäjien esittämä väite, jonka mukaan kehykset voidaan jakaa *eksplisiittisiin* ja *implisiittisiin* kehyksiin. Eksplisiittisellä kehyksellä tarkoitetaan suoraan sanaston avulla esitettävää kehystä, kun taas implisiittinen kehys perustuu keskustelun piileviin ulottuvuuksiin, kuten sanojen esiintymiseen toistensa yhteydessä. Semanttinen kartta on tarkoitettu implisiittisten kehysten analysointiin.

Semanttinen kartta perustuu Millerin menetelmän tavoin sanojen yhdessä esiintyvyyden analysointiin kosinin avulla. Semanttisen kartan muodostamisen ensimmäinen vaihe on siis kosinin soveltaminen kaikille dokumenttijoukon sanoille samaan tapaan kuin Millerin menetelmässä.

Kartan muodostamisen toinen vaihe on kosinimatriisin visualisointi verkkona siten, että jokainen sana on verkon solmu [Hellsten et al., 2010]. Hellsten, Dawson ja Leydesdorf generoivat visualisaation työkalulla, joka muodosti visualisaation muokkaamalla solmujen paikkaa iteratiivisesti 'jyrkimmän laskun' (steepest descent) -menetelmällä [Debye, 1909]. Jyrkimmän laskun menetelmä optimoi solmujen paikkaa muuttamalla jokaisen solmun paikkaa aina siihen suuntaan, joka parantaa eniten solmujen etäisyyksien vastaavuutta kosinitaulukkoon.

4.3.2 Menetelmien sovelluskohteet

Laskennallisen kehysanalyysin sovelluskohteiksi on kaikissa esimerkkitutkimuksissa otettu tiettyä, valittua aihetta koskeva mediakeskustelu. Aiheen valinnan avulla voidaan tutkia kehyksiä, joissa tiettyä aihetta koskeva keskustelu esitetään. Joissakin laskennallista kehysanalyysiä soveltaneissa tutkimuksissa aihe on rajattu avainsanojen avulla [Hellsten et al., 2010], toisissa taas datalähteen ja ajanjakson avulla [Pashakhin, 2016].

Miller käytti menetelmäänsä kosteikkoihin liittyvän mediakeskustelun analysointiin. Valintaansa hän perusteli sillä, että keskustelussa oli selkeästi havaittavissa eri osapuolia, kuten luonnonsuojelijat ja kiinteistönomistajat. Datana hän käytti 102 kosteikkokeskusteluun liittyvää lehtiartikkelia, joissa mainittiin tai jotka olivat peräisin eri sidosryhmiltä. Aluksi jokaiselle artikkelille annettiin sidosryhmäkoodi myöhempää sidosryhmäkohtaista analyysiä varten.

Miller otti menetelmänsä vaiheen 2 jälkeen tarkempaan analyysiin 123 usein esiintyvää sanaa. Tämän jälkeen sanoille ja niiden frekvensseille sovellettiin menetelmän myöhempiä vaiheita. Löydettyjä kehyksiä olivat esimerkiksi tulvan ja vesistöjen kehykset [Miller, 1997, s. 372]. Vesistöjen kehyksessä kehyksen määrittäviä avainsanoja olivat esimerkiksi vesi, globaali, tuho ja rakentaminen.

Pashakhin [Pashakhin, 2016] sovelsi Latent Dirichlet Allocation -aihemallinnusta Ukrainan kriisin (vuosina 2013-2014) mediakeskustelun analysointiin. Hän

TABLE 1
News Release Key Term Cluster Groupings and Code Source Identifier Terms

	Eigenvector Values			
Cluster Name	*1*	*2*	*3*	*Terms in Cluster*
(1) Habitat	−.37	−.34	−.39	Habitat, bird, migratory, birds, rare
(2) Endangered	−.38	−.10	−.24	Endangered, protect, acres, everglades, restored, threatened, project, fish
(3) Wildlife	−.52	.13	−.28	Wildlife, natural, earth
(4) Conservation	−.45	−.02	.04	Conservation, watershed, ecosystems
(5) Environmental	−.22	.51	−.20	Environmental, plant, plants, field, quality
(6) Waters	−.26	.65	−.00	Waters, global, damage, construction, projects
(7) Flood	−.15	−.00	.77	Flood, homes, flooding, management, floodplain, future, levees, corps, engineers, runoff, resources, control
(8) Farm	.68	.07	−.12	Farm, state, legislation, agencies, supports, issue, farmers, property, compensation, owners, agriculture, ranchers, congressional, rights, support, legislative
(9) Landowners	.60	−.04	−.19	Landowners, statewide, lawmakers, court, balance, agricultural, definition
(10) Regulation	.16	.57	.07	Regulation, act, public, decision, impact, economy, clean
(11) Policy	.38	.32	−.05	Policy, land, government, issues, regulatory, economic, health
(12) Conservation advocates	−.52	.08	.0	[Source Term]
(13) Property-owner advocates	.72	.05	−.15	[Source Term]

Kuva 2: Millerin kosteikkokeskustelun analyysin tuloksena olleet sanaklusterit [Miller, 1997]

perusteli valintaansa muun muuassa sillä, että aihemallinnus ei vaadi tutkijalta aiempaa tietämystä aiheesta tai ohjattua luokittelua. Datana hän käytti televisiolähetyslitterointeja kahdelta eri kanavalta: Venäjän TV 1:ltä sekä Ukrainan kanava 5:ltä [Pashakhin, 2016]. Pashakhin rajasi datansa siten että se sisälsi ajallisesti kaikki Ukrainan kriisin päätapahtumat, siihen kuuluen rajauksen jälkeen noin 25 000 tekstiä Venäjän TV 1:ltä ja noin 20 000 tekstiä Ukrainan kanava 5:ltä. Ukrainan kanava 5:n tekstit käännettiin automaattisella käännösohjelmalla venäjän kielelle, jotta dataa molemmista lähteistä pystyttiin käyttämään samassa mallissa.

Pashakhin sovelsi aihemallinnusta käyttäen parametrien estimointiin Gibbs -otantaa [Pashakhin, 2016, Geman ja Geman, 1984], ja valiten ennaltamääritetyksi aihemääräksi 100 aihetta. Koska LDA on luonteeltaan stokastinen eli sisältää satunnaisuutta, Pashakhin otti myös aiheiden vaihtelevuuden eri ajokerroilla huomioon tutkimuksessaan [Pashakhin, 2016]. Hän ajoi mallinnuksen 5 kertaa, generoiden jokaiselle aiheparille samankaltaisuusarvon [Pashakhin, 2016, Koltsov et al., 2014]. Mikäli kaksi aihetta olivat

riittävän samanlaisia keskenään, ne määriteltiin samaksi aiheeksi. Tämän jälkeen lopullisiksi, stabiileiksi aiheiksi hyväksyttiin aiheet, jotka esiintyivät vähintään kolmella ajokerralla viidestä.

Aihemallinnuksen tuloksena oli 49 stabiilia aihetta, jotka nimettiin käsin [Pashakhin, 2016]. Tulosten mukaan aiheet erosivat kahden eri televisiokanavan lähetysten välillä. Osa lopputuloksena olleista aiheista ei liittynyt Ukrainan kriisiin (esimerkiksi aihe "IT News"), mutta kriisiin liittyneiden aiheiden avulla pystyttiin muodostamaan kuva kahden eri kanavan kehystämiseroista. Tulosten mukaan Venäjän TV 1 käytti raportoinnissaan enemmän sotaan liittyviä sanoja aiheita, kun taas Ukrainan kanava 5 kehysti kriisin sarjana terroristisia rikoksia [Pashakhin, 2016].

Hellsten, Dawson ja Leydesdorff sovelsivat semanttisten karttojen kehysanalyysiä New York Timesissa käydyn keinotekoisten makeutusaineiden mediakeskustelun vertailuun 1980-luvulla ja 2000-luvulla [Hellsten et al., 2010]. He käyttivät datana The New York Timesin artikkeleita vuosilta 1984-1986 (16 artikkelia) sekä 2004-2006 (8 artikkelia). He generoivat 1980-luvulla ilmestyneistä uutisista oman semanttisen karttansa ensin artikkeleille jotka sisälsivät sanan "artificial sweetener" ja tämän jälkeen omat karttansa kahdelle kartassa esiintyneelle makeutusaineelle. Tämän jälkeen he generoivat samoista aiheista kartat aineistonaan 2000-luvulla ilmestyneitä uutisia.

Analyysin tuloksissa havaittiin eroja keskustelun kehyksissä eri vuosikymmenillä. Esimerkiksi 2000-luvulla aspartaamikirjoittelussa esiintyi toisiinsa yhteyksissä tiettyjä tuotemerkkejä, kuten Pepsi, Coca-Cola ja Splenda, joita ei keskustelussa 1980-luvulla esiintynyt. Lisäksi esimerkiksi 1980-luvulla sana syöpä oli kartassa hyvin lähellä sanaa aspartaami, kun 2000-luvun kartassa sen esiintyvyydet eivät ylittäneet visualisaation rajaa lainkaan [Hellsten et al., 2010].

4.3.3 Laskennallisten kehysanalyysimenetelmien kritiikki

Laskennallisten kehysanalyysimenetelmien kehittäminen ja soveltaminen ei ole ongelmatonta. Seuraavaksi esitellään kaksi yleistä laskennallisia kehysanalyysimenetelmiä kohtaan esitettyä kritiikkiä: kehysten analysoiminen aiheiden kaltaisina, sekä kehysten limittyminen dokumenteissa. Tämän jälkeen esitellään erityisesti Latent Dirichlet Allocation -aihemallinnuksen käyttöön kehysanalyysissä liittyvää kritiikkiä.

Luvun 4.1 kehyksen määritelmien pohjalta voidaan sanoa, että kehyksen pitäisi kuvailla merkityksiä joita tietylle aiheelle annetaan, ei aihetta itseään. Edellisessä aliluvussa esitellyistä menetelmistä Latent Dirichlet Allocation -aihemallinnuksen nimikin kertoo menetelmän olevan kehitetty aiheiden, ei kehysten tunnistamiseen. Jo Milleriä kritisoitiin kehysanalyysitutkimuksistaan siitä, että lopputulokset ovat enemmänkin aiheita kuin kehyksiä [Carragee ja Roefs, 2004]. Tämä on havaittavissa myös kosteikkotutkimuksesta: kehyksen pitäisi kuvailla merkityksiä joita tietylle aiheelle annetaan, kun taas eräät

esimerkkikehykset tulva ja vesistö kuvaavat selvästi vain aiheita [Carragee ja Roefs, 2004].

Toisaalta taas esimerkiksi Entmanin kehystämisen määritelmä (luku 4.1), ”joidenkin näkökantojen valitseminen ja tekeminen merkittävämmäksi tekstillä kommunikoiden” voisi hyväksyä myös aiheet kehyksiksi, mikäli aiheet voidaan katsoa näkökannoiksi. Gamsonin ja Laschin tulkintapakettien määritelmän mukaan kehystäminen voi myös sisältää perustelua, mikä taas on ristiriidassa kehysten aiheiksi operationalisoimisen kanssa. Kaiken kaikkiaan siis kehystämisen eri määritelmät sallivat enemmän tai vähemmän kehyksen operationalisoimisen aiheeksi, ja kehystämisen laskennallinen operationalisointi on aina tulkinnanvaraista.

Toinen laskennalliseen kehysanalyysiin liittyvä keskeinen ongelma on kehysten limittyminen. LDA:ssa sanoja käsitellään järjestyksettömänä joukkona, ja lopputuloksena, kuten kuvasta 3 nähdään, aiheet voivat jakautua sana kerrallaan dokumentin alueelle. Kuvassa 3 nähtävä aiheiden eli kehysten limittyminen LDA:ssa on ristiriidassa esimerkiksi Davidin, Atunin, Fillen, ja Monterolan käyttämien kehysanalyysimenetelmien oletuksen, että yksi dokumentti sisältää vain yhden kehyksen, kanssa. Toisaalta taas luvun 4.1 kehysanalyysin määritelmissä ei spesifioida, kuinka monta kehystä yhdessä tekstissä voi olla. Kuitenkin, mikäli yhdessä artikkelissa voisi olla monta kehystä, on kyseenalaista, voivatko kehykset koskaan limittyä esimerkiksi kuvan 3 kaltaisesti niin, että eri kehysten sanat on hajautettu koko artikkelin laajuudelle.

Myös semanttisiin karttoihin liittyy kehysten dokumenttikohtaisen limittymisen ongelma. Semanttisissa kartoissa sanojen esiintyvyydet lasketaan dokumenttikohtaisesti, joten mikäli esimerkiksi yhdessä dokumentissa esiintyy usein kaksi kehystä toistensa kanssa, kartassa molempiin kehyksiin liittyvät sanat ilmaantuvat lähelle toisiaan. Kehysten tunnistaminen kartan perusteella ei siis voi perustua pelkästään klustereihin, vaan vaatii myös ihmisen arviointia. Myös semanttisiin karttoihin liittyy kysymys siitä, ovatko tulokset aiheita vai kehyksiä. Esimerkiksi Hellsten, Dawson ja Leydesdorff itse huomauttavat, että semanttiset kartat eivät tunnista kehyksiin liittyviä tunteellisia latauksia [Hellsten et al., 2010].

Latent Dirichlet Allocation -aihemallinnukseen liittyy omat erityiset kritiikin aiheensa. Koska lopputuloksena olevat aiheet ovat vain sanajoukkoja, yleensä ihminen nimeää aiheet joukkojen perusteella, kuten kuvan 3 esimerkissä on tehty. Huomioonotettavat seikat voidaan jakaa sekä itse malliin liittyviin, että sen käyttöön kehysanalyysissä liittyviin ongelmiin. Esimerkiksi itse mallissa aiheiden sisältö vaihtelee valitun aiheiden määrän mukaan, eikä aihemallinnuksessa ole yhtä yleisesti hyväksyttyä aiheiden määrää. Mikäli aiheita on vähän, aiheet saattavat sekoittua keskenään, ja mikäli aiheita on paljon, aiheisiin saattaa päätyä monia samankaltaisia aiheita [Greene et al., 2014].

LDA-aihemallinnuksen arviointiin on kehitetty sekä automaattisia että

Kuva 3: Blein, Ng:n, ja Jordanin esimerkkikuva LDA-aihemallinnuksen tuloksesta. Yllä olevat sanajoukot ovat aiheita, jotka ihminen on nimennyt otsikoilla. Eri aiheisiin kuuluvat sanat on väritetty eri väreillä dokumentissa. [Blei et al., 2003]

manuaalisia menetelmiä, joita voidaan käyttää esimerkiksi sopivan aihemäärän määrittämiseen. Menetelmät voidaan jakaa mallin ennustavuutta sekä mallin tulkintaa mittaaviin menetelmiin. Esimerkiksi Greene, O'Callaghan ja Cunningham kehittivät automaattisen menetelmän, joka arvioi mallin ennustavuutta. Menetelmä arvioi sopivaa aiheiden määrää sen perusteella, kuinka stabiileja aiheet ovat aineiston kokoa ja siihen kuuluvia dokumentteja vaihdeltaessa [Greene et al., 2014]. Menetelmässä luodaan ensin aihemallinnus koko aineistosta, ja tämän jälkeen tietyn kokoisista osajoukoista aineistoa. Tämän jälkeen osajoukkojen mallinnuksesta syntyneitä aiheita verrataan yksitellen koko aineiston mallinnuksen aiheisiin. Jokaiselle mallinnusparille lasketaan eräänlainen tunnusluku siitä, kuinka samankaltaisia aiheet ovat eri aineistojoukkojen välillä. Lopuksi näistä tunnusluvuista lasketaan keskiarvo, joka kuvaa sitä, kuinka stabiili aihemallinnus on kyseisellä aiheiden määrällä. Tämän jälkeen samanlainen stabiiliuden arviointi tehdään jokaiselle aihemäärälle, ja arviointeja vertaillaan keskenään [Greene et al., 2014].

Todennäköisyyksiin perustuvia aihemallinnuksen ennustettavuutta mittaavia arviointimenetelmiä on kritisoitu siitä, että ne eivät mittaa itse mallin tuloksen sisäistä esittävyyttä [Chang et al., 2009]. Edellisessä luvussa esi-

telty menetelmä voisi tuottaa korkean stabiliteetin aihemallille, vaikka itse aiheet eivät kuvaisi lainkaan aineistoa tai aiheet eivät olisi sisäisesti yhtenäisiä. Tämän vuoksi esimerkiksi Chang, Boyd-Graber, Gerrish, Wang ja Blei ovatkin kehittäneet manuaalisia tapoja mitata aihemallin tulkittavuutta [Chang et al., 2009]. Ensimmäisessä heidän kehittämässä menetelmässä ihmisen tuli tunnistaa tietyn aiheen sanojen joukosta ”ylimääräinen” sana, joka kuului pienellä todennäköisyydellä kyseiseen aiheeseen mutta suurella toiseen aiheeseen. Toisessa menetelmässä taas ihmisen tuli otsikon ja muutaman ensimmäisen lauseen perusteella tunnistaa dokumentista, mikä annetuista aiheista kuvaa dokumenttia huonoiten [Chang et al., 2009]. Näihin menetelmiin tehtyjen koeasetelmien tuloksista muodostettiin tunnuslukuja kuvaamaan ihmisen ja koneen välistä yhteisymmärrystä.

5 Ohjatun koneoppimisen soveltaminen tekstianalyysiin

Tässä luvussa esitellään ohjatun koneoppimisen teoriaa tekstianalyysin näkökulmasta. Ensimmäisessä aliluvussa esitellään ohjatun koneoppimisen soveltamisen yleinen prosessi, jonka jälkeen toisessa aliluvussa esitellään datankeruuseen liittyviä kysymyksiä. Kolmannessa aliluvussa esitellään piirteiden eristämisen perusperiaatteet, ja neljännessä aliluvussa alkioiden opetus- ja testijoukkoon jakamisen periaatteita. Viidennessä aliluvussa käsitellään ratkaisuja epätasaiseen opetusdatan luokkajakaumaan. Lopuksi esitellään koneoppimismallin valintaan liittyviä seikkoja sekä kaksi yleistä koneoppimismallia, satunnaismetsäluokittelija ja tukivektorikone.

5.1 Ohjatun koneoppimisen perusperiaatteet

5.1.1 Formalisointi

Koneoppimismenetelmät sopivat Kotsiantisin [Kotsiantis, 2007] mukaan erityisen hyvin tilanteisiin, jossa ihmisen on vaikea tunnistaa yhteyksiä erilaisten datan ominaisuuksien välillä. Koneoppimisalgoritmien perusperiaate on, että jokainen datan dokumentti esitetään saman piirrejoukon (feature set) avulla. Piirteet voivat olla niin jatkuvia, kategorisia kuin binäärisiäkin. Piirteiden lisäksi jokaista datan dokumenttia vastaa jonkinlainen *vaste*, joka voi olla myöskin jatkuva, binäärinen tai kategorinen. Esimerkiksi kuvan 4 ongelmassa jokaista dokumenttia vastaa binäärinen vaste (kuvan 4 kolumni ”class”).

Ohjatuksi koneoppimiseksi kutsutaan tilannetta, jossa saatavilla on dataa jolle oikeiden vasteiden arvot tiedetään. Vastemuuttujia kutsutaan myös *luokiksi*, kun kyseessä on kategorinen tai binäärinen luokittelu. Ohjaamattomaksi koneoppimiseksi kutsutaan tilannetta, joissa oikeita luokkia ei tiedetä, vaan mallinnuksen tarkoituksena on selvittää dokumenteille oikeat luokat tai vastemuuttujat.

Ohjatun koneoppimisen perusperiaatteisiin kuuluu, että mallin halutaan suoriutuvan mahdollisimman hyvin uuden datan luokittelusta. Mallin suorituskykyä arvioidaan usein ohjatussa koneoppimisessa *yleistysvirheen* estimoinnilla [Shalev-Shwartz ja Ben-David, 2014, s. 34]. Yleistysvirheellä tarkoitetaan esimerkiksi kategorisen luokittelijan tapauksessa todennäköisyyttä, että luokittelija tekee väärän luokituksen uudelle datalle, joka on samasta jakaumasta kuin sen opetuksessa käytetty data.

Formaalimmin ilmaistuna, data jolle luokat tiedetään ja jota käytetään mallin opetuksessa, koostuu pareista $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$, missä x ilmaisee tämän tutkielman tapauksessa vastamedia-artikkelia tai siitä eristettyjä piirteitä, ja y kehystämisen prosessin kategoriaa, jotka esitellään myöhemmin luvussa 6. Laskennallisesti ongelma voidaan muotoilla funktion $\hat{f}$ estimoimiseksi tunnettujen datapisteiden avulla siten, että jokaiselle uudelle alkiolle (x_0, y_0), missä y_0 on tuntematon, $\hat{f}(x_0)$ halutaan olevan mahdollisimman lähellä y_0:aa [James et al., 2014, s. 21].

Luokitteluongelmissa yleistysvirheen estimointia kutsutaan usein myös luokitteluvirheen estimoinniksi. Kun jokainen virhe luokittelussa on yhtä merkittävä, luokitteluvirhe lasketaan väärien luokitusten osuutena kaikista luokituksista. Kun luokitellaan n datapistettä, luokitteluvirhe määritellään siis seuraavana keskiarvona [James et al., 2014, s. 37]:

$$L = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{f}(x_i)), \tag{4}$$

missä indikaattorifunktion arvo määräytyy sen mukaan, onko mallin tekemä luokitus sama kuin datapisteen oikea luokka. Tällöin luokittelutarkkuudeksi kutsutaan luokitteluvirheen komplementtia eli oikeiden luokitusten osuutta kaikista luokituksista.

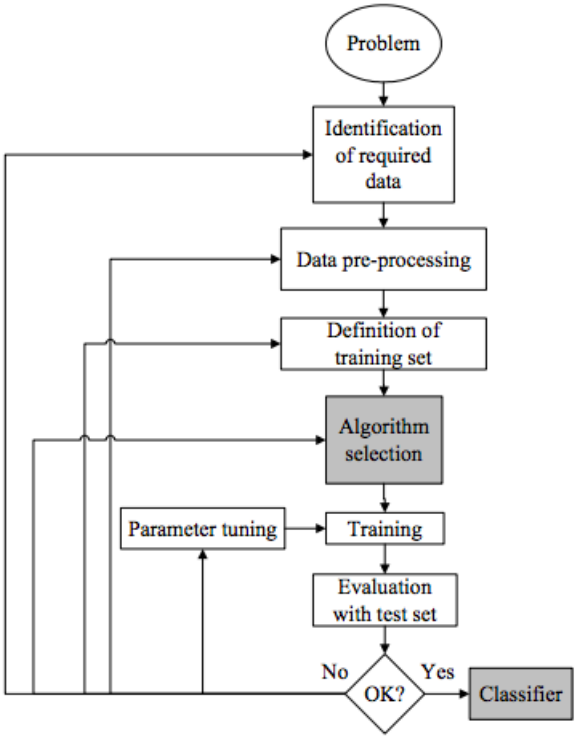
Data in standard format					
case	Feature 1	Feature 2	...	Feature n	Class
1	xxx	x		xx	good
2	xxx	x		xx	good
3	xxx	x		xx	bad
...					...

Kuva 4: Ohjattuun koneoppimiseen soveltuvan aineiston dokumentit tiedetyillä vastemuuttujilla [Kotsiantis, 2007]

5.1.2 Ohjatun koneoppimisen soveltamisen vaiheet

Ohjatun koneoppimisen soveltamiseen käytännön ongelmaan koostuu Kotsiantisin [Kotsiantis, 2007] mukaan erilaisista vaiheista, joiden välillä siirrytään tiettyjen ehtojen mukaan (kuva 5). Hänen mukaansa mallinnus alkaa

ongelmasta johon ratkaisua etsitään. Tämän jälkeen identifioidaan sopiva data ja tehdään tarvittava esiprosessointi. Datan esiprosessoinnin jälkeen määritellään opetusdata, jonka jälkeen tulee valita itse koneoppimismalli. Kun sopiva koneoppimismalli on valittu, iteroidaan mallin opetusta, evaluointia sekä parametrien asetusta kunnes parhaat parametrit kyseiselle mallille on valittu. Näiden vaiheiden jälkeen voidaan edelleen palata lopusta opetusdatan määrittelyyn, datan esiprosessointiin tai jopa itse datan valintaan.



Kuva 5: Ohjatun koneoppimisen soveltaminen käytännön ongelmaan [Kotsiantis, 2007]

5.2 Datankeruu ja esiprosessointi

5.2.1 Datankeruu ja sen haasteet

Ohjatun koneoppimisen ensimmäinen askel on datankeruu. Datankeruu on yleinen pullonkaula ohjatussa koneoppimisessa, sillä valmiiksi luokiteltua dataa jolle lopputulokset tiedetään, on usein hankalaa löytää riittävästi. Naiivi lähestymistapa datankeruuseen on luokitella käsin riittävästi opetusdataa. Riittävän datamäärän käsin luokittelu on kuitenkin hyvin työlästä, ja luokiteltu data soveltuu usein vain hyvin erikoistuneisiin tarkoituksiin [Roh et al., 2018].

Eräitä ratkaisutapoja vähäisen luokitellun datan ongelmaan ovat aktiivinen oppiminen (active learning) sekä joukkoistamalla datan kerääminen [Roh et al., 2018, Settles ja Craven, 2008]. Seuraavaksi esitellään näiden lähestymistapojen perusperiaatteet.

Aktiivisen oppimisen perusajatuksena on, että luokiteltua dataa hankitaan vähitellen mallin opetuksen tarpeisiin. Aktiivisessa oppimisessa alussa valitaan luokiteltu pieni perusdata, jonka jälkeen jollain hakukriteerillä lisätään luokiteltuja alkioita perusjoukkoon ja opetetaan malli uudestaan. Uusien luokiteltujen alkioiden lisäämisellä halutaan saavuttaa maksimaalinen luokittelijan tarkkuuden kasvu: sen vuoksi hakuun on kehitetty useita erilaisia lähestymistapoja [Settles ja Craven, 2008].

Eräitä lähestymistapoja uusien luokitusten hankkimiseen aktiivisessa oppimisessa ovat *epävarmuusotanta* (uncertainty sampling) sekä *kyselykomitea* (Query-By-Committee). Epävarmuusotannassa hakukriteerinä seuraavan alkion valinnassa on valita alkio, jonka luokittelu senhetkisellä mallilla tuottaa epävarmimman tuloksen. Hakukriteeri voidaan siis formalisoida x:n etsinnäksi, jolle seuraava funktio on maksimoitu:

$$f(x) = 1 - P(y^*|x), \tag{5}$$

missä y^* on luokittelijan antama todennäköisin luokka alkiolle x. Kyselykomitealla taas tarkoitetaan usean mallin opetusta, ja tämän jälkeen uuden luokiteltavan alkion valitsemista sillä perusteella, minkä alkion luokasta eri mallit ovat erimielisimpiä. Jokainen kyselykomitean malli opetetaan eri perusjoukosta valitulla opetusdatasetillä [Settles ja Craven, 2008].

Aktiivinen oppiminen sopii hyvin tekstianalyysiin, jossa usein saatavilla on paljon luokittelematonta dataa mutta datan luokittelu on aikaavievää [Settles ja Craven, 2008]. Aktiivinen oppiminen ei kuitenkaan ota kantaa siihen, millä tavalla luokiteltua dataa käytännössä hankitaan. Eräs ratkaisu luokitellun datan nopeampaan ja halvempaan hankintaan on *joukkoistaminen* (crowdsourcing) [Lease, 2011]. Joukkoistamisella tarkoitetaan datan luokittelun ulkoistamista nimetyiltä toimijoilta ennalta määrittelemättömälle, usein suurelle joukolle ihmisiä [Quinn ja Bederson, 2011].

Luokitellun datan keräämiseen joukkoistamalla liittyy erilaisia huomioonotettavia asioita, kuin tilanteeseen jossa opetusdatan luokittelu tehdään tutkijan ja hänen kollegoidensa kesken. Molemmille tavoille yhteistä on, että mikäli luokittelijoita on useampi kuin yksi, tarvitaan jonkunlaisia mittareita luokittelijoiden välisen samanmielisyyden (inter-rater reliability tai inter-rater agreement) arviointiin [Lease, 2011].

Luokittelijoiden samanmielisyyden arviointiin käytetään usein Kappa-testiä [Cohen, 1960]. Kappa-testi olettaa, että molempien luokittelijoiden mielipiteet ovat samanarvoiset, ja laskee luokitteluista Kappa-arvon kaavalla

$$K = \frac{p_0 - p_c}{1 - p_c}, \tag{6}$$

missä p_0 on samojen luokitusten osuus kaikista luokituksista ja p_c on todennäköisyys, jolla luokittelijat sattumalta tekivät keskenään saman luokituksen. Erotus $p_0 - p_c$ kuvastaa niiden luokitusten osuutta, joissa sama luokitus ei tapahtunut sattumalta.

Opetusdatan luokittelun jakamisessa useammalle ihmiselle on tärkeää, että luokittelijoilla on mahdollisimman samankaltainen käsitys luokittelusäännöistä. Useissa tapauksissa luokittelu on kuitenkin hyvin tulkinnanvaraista ja luokittelusäännöt elävät vielä opetusdatan luokittelun alettua, sillä joskus hyvin vaikeaa muodostaa heti kaikki datan erikoistapaukset kattavaa luokittelusäännöstöä [Settles ja Craven, 2008]. Joukkoistamisen tapauksessa erityisongelmaksi tulee, miten luokittelijat voivat kommunikoida toistensa kanssa luokittelusäännöistä ja luokitteluiden yhteneväisyydestä.

Luokittelijoiden toisistaan eristäytyneisyyden haittoja voidaan kompensoida huolehtimalla, että jokainen luokiteltava alkio saa useamman luokituksen, ja valitsemalla tämän jälkeen lopulliseksi luokitukseksi esimerkiksi eniten kannatusta saanut luokka [Settles ja Craven, 2008]. Luokittelijoita voidaan myös arvioida automaattisesti testaamalla, miten he suoriutuvat sellaisen aineiston luokittelusta, jolle oikeat luokitukset tiedetään [Quinn ja Bederson, 2011].

5.2.2 Tekstidatan esiprosessointi

Esiprosessoinnin tarkoituksena on erottaa raakadatasta relevantti tieto epärelevantista. Esiprosessointi tehdään koneoppimisen prosessissa ennen piirteiden eristämistä. Seuraavaksi esitellään yleisimpiä luonnollisen kielen käsittelyssä käytettäviä datan esiprosessointimenetelmiä.

Stemmauksella (stemming) tarkoitetaan sanojen perusmuotoiseen juuriesitykseen muuttamista [Kannan ja Gurusamy, 2015]. Stemmausta pidetään hyödyllisenä varsinkin, jos tarkoitus on laskea sanojen esiintyvyysmääriä [Lovins, 1968]. Tällöin esimerkiksi erilaiset sija- ja omistuspäätteet eivät vaikuta lopputulokseen, jos stemmaus toimii oikein ja muuttaa saman sanan erilaiset muodot samaksi juureksi.

Eräs yleinen stemmausalgoritmi on Porter-stemmaus [Porter, 1997], joka perustuu ajatukselle siitä, että englannin kielessä monimutkaisemmat sanapäätteet koostuvat yksinkertaisemmista päätteistä. Porter-stemmauksessa on kuusi vaihetta, joista jokaisessa sanapäätteitä korvataan toisilla tiettyjen sääntöjen mukaan. Esimerkiksi ensimmäisessä vaiheessa päätteet ”SSES” korvataan päätteillä ”SS”.

Toinen yleinen esiprosessointitapa luonnolliselle kielelle on pysäytyssanojen poistaminen. Pysäytyssanoilla tarkoitetaan sanoja, joita voi esiintyä tekstissä paljon, mutta joiden merkitys analyysille on pieni [Wilbur ja Sirotkin, 1992]. Tällaisia sanoja ovat suomen kielessä esimerkiksi *ja*, *vähemmän* ja *sinä*. Yleinen keino pysäytyssanojen poistamiseen on valmiiden pysäytyssanalistojen käyttäminen.

5.3 Piirteiden eristäminen

Datan esiprosessoinnin jälkeen tulee päättää, millaisia piirteitä koneoppimisalgorimia varten halutaan eristää datasta. Yleisenä ohjeena piirteiden eristämisessä pidetään ”brute-force” -tekniikkaa, eli kaikkien [Kotsiantis, 2007] saatavilla olevien ja mahdollisesti hyödyllisiksi kuviteltavissa olevien piirteiden eristämistä. Brute force -tekniikassa on kuitenkin myös omat huonot puolensa, sillä se voi tuottaa paljon myös luokittelun kannalta hyödytöntä piirredataa.

Eräs yleinen tekstidatasta eristettävä, monelle koneoppimisalgoritmille hyödyllinen piirre on eri sanojen esiintyvyydet eri dokumenteissa. Tätä piirrettä kutsutaan myös lyhenteellä TF (term frequency). Pelkkien sanojen esiintymismäärien käyttäminen piirteinä aiheuttaa kuitenkin myös hyödyttömän informaation keräämistä: esimerkiksi sana ”the” esiintyy usein englanninkielisissä dokumenteissa, mutta koska se esiintyy usein *kaikissa* dokumenteissa, esiintymismäärän kerääminen vääristää muiden sanojen esiintymismääristä saatavaa informaatiota.

Edellä kuvattuun ongelmaan on olemassa ratkaisu, joka ottaa huomioon myös sanojen esiintyvyyden koko korpuksessa. Sitä kutsutaan termillä IDF (inverse document frequency), ja se määräytyy sen mukaan, kuinka monessa dokumentissa sana esiintyy koko korpuksessa [Jones, 1972, Robertson, 2004]. IDF määräytyy kaavalla:

$$IDF = \log \frac{N}{n_i}, \tag{7}$$

missä N on koko aineiston dokumenttien määrä ja n_i on niiden dokumenttien määrä, joissa sana esiintyy [Robertson, 2004]. IDF-kerroin on siis sitä pienempi, mitä useammassa dokumentissa sana esiintyy. Kun sanan esiintyvyysmäärä TF kerrotaan sanan IDF-termillä, saadaan informaation hyödyllisyyden kannalta pelkkää sanan esiintyvyysmäärää huomattavasti luotettavampi mittari TF-IDF. [Robertson, 2004].

Toinen yleinen luonnollisen kielen käsittelyssä käytettävä piirre on *sanavektorit* (dense embeddings, word embeddings) [Goldberg, 2017]. Sanavektoreiden ideana on esittää jokainen sana d -ulotteisena vektorina, missä d on huomattavasti pienempi kuin koko sanaston koko. Esimerkiksi jos koko sanaston koko on 40 000 sanaa, sanavektoritaulukko voidaan luoda vain 100-ulotteiseksi [Goldberg, 2017, s. 90]. Sanavektoreiden luomisessa perusperiaatteena on, että samankaltaiset sanat halutaan esittää samankaltaisten vektoreiden avulla. Samankaltaisuus voidaan määritellä esimerkiksi siten, että sanat esiintyvät samankaltaisessa kontekstissa. Sanavektoreiden luomiseen on kehitetty useita erilaisia algoritmeja, kuten esimerkiksi kontekstimatriisien ulotteisuuden pienentäminen (dimensionality reduction) [Goldberg, 2017, s. 121], joita ei käsitellä tässä tutkielmassa enempää.

Tietyt neuroverkkoalgoritmit mahdollistavat myös ohjatun koneoppimisen,

jossa erillistä tutkijan tekemää piirteiden eristämistä ei tarvita. Yksi esimerkki näistä on konvoluutiota hyödyntävä neuroverkko [Goodfellow et al., 2016, s. 326]. Konvoluutiota hyödyntävä neuroverkko on neuroverkko, jossa ainakin yhdessä tasossa käytetään matriisitulon sijaan konvoluutiota. Konvoluutio on alun perin matematiikasta lähtöisin oleva käsite, joka määritellään matriiseille I (syöte) ja K (kernel) seuraavasti:

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n) * K(i-m, j-n) \tag{8}$$

Koska konvoluutioneuroverkoissa K on usein pienempi kuin syöte, konvoluution avulla voidaan tunnistaa tuhansia pikseleitä sisältävistä kuvista pieniä elementtejä kuten reunoja [Goodfellow et al., 2016, s. 330]. Konvoluutiota hyödyntäviä neuroverkkoja on käytetty useiden luonnolliseen kieleen liittyvien ongelmien ratkaisussa. Esimerkiksi Hu, Lu, Li ja Chen mallinsivat lauseita konvoluutiota hyödyntävällä neuroverkolla [Hu et al., 2014]. He esittivät lauseen sanoja vastaavien sanavektorien avulla ja altistivat tämän jälkeen sanavektorisarjat konvoluutiolle. He hyödynsivät lähestymistapaansa lauseiden yhdistämisen ongelmaan, testaten malliaan esimerkiksi twiittien ja niihin tulleiden vastauksen yhdistämisessä.

5.4 Yleistysvirheen estimointi alkioiden jakamisella

Luokitellun datan jakaminen joukkoon jolla malli opetetaan, ja joukkoon jolla yleistysvirhettä estimoidaan, on yleinen tapa arvioida mallin suorituskykyä. Näitä joukkoja voidaan kutsua esimerkiksi opetus- ja testijoukoiksi. Datan jakamiseen opetus- ja testijoukkoihin on olemassa monia erilaisia periaatteita, jotka koskevat sekä joukkojen kokojen valintaa, että tapaa jolla joukkojen alkiot valitaan datasta.

Alkioiden valinta voidaan tehdä joko satunnaisesti tai *rationaalisella* valinnalla [Golbraikh ja Tropsha, 2000]. Rationaalisessa valinnassa tavoitteena on, sekä opetus- että testijoukon jakaumat muistuttavat alkuperäisen datan jakaumaa, eli esimerkiksi että jokainen testijoukon piste on lähellä ainakin yhtä opetusjoukon pistettä [Golbraikh ja Tropsha, 2000]. Esimerkiksi K-means-klusterointia on käytetty rationaalisessa valinnassa siten, että testijoukon alkiot valitaan tasaisesti eri klustereista.

Opetus- ja testijoukkojen suuruuden valitsemiseen ei ole olemassa yhtä oikeaa tapaa. Hyödyllinen tapa tutustua opetus- ja testijoukon merkitykseen on esittää koneoppimismallin virhe biasin ja varianssin avulla. Tiedetään, että neliöidyn virheen odotusarvo voidaan esittää kaavan

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0))]^2 + Var(\epsilon) \tag{9}$$

avulla, missä $Var(\hat{f}(x_0))$ on funktion varianssi, $[Bias(\hat{f}(x_0))]^2$ on neliöity *biasin* aiheuttama virhe, eli siis virhe joka aiheutuu ongelman yksinkertaistamisesta malliksi, ja $Var(\epsilon)$ on virhe jonka alkuperää ei voida määrittää,

esimerkiksi satunnaisvirhe [James et al., 2014]. Tavoitteena on, että kaikki virheen komponentit voitaisiin minimoida.

Funktiolla, joka on sovitettu tarkasti opetusdataan, on yleensä suurempi varianssi, mutta pienempi bias kuin funktiolla, jossa ollaan tehty yleistyksiä opetusdatasta [James et al., 2014]. Opetus- ja testijoukon valintaa voidaankin ajatella biasin ja varianssin kautta: kun opetusjoukon kokoa kasvatetaan, mallin varianssia on mahdollista pienentää. Toisaalta jos opetusjoukon kokoa kasvatettaessa testijoukon koon tulee pienentyä, tällöin jää vähemmän mahdollisuuksia testata mallin todellista biasta sekä varianssin vaikutusta tuloksiin. Yleisesti käytettyjä opetus- ja testijoukkojen jakamiskäytäntöjä ovat datan jakaminen siten, että 90% datasta kuuluu opetusjoukkoon ja 10% testijoukkoon [Breiman, 2001], sekä siten että 80% datasta kuuluu opetusjoukkoon ja 20% testijoukkoon [Lang, 1995].

Luokitellun datan alkioiden jakaminen kerran opetus- ja testijoukkoon antaa tavan vertailla erilaisia malleja, mutta ei anna luotettavaa estimaattia mallin virheen suuruudesta. Opetus- ja testijoukkojen jaon ollessa satunnaisuuteen perustuva, myös virhe-estimaatti vaihtelee testijoukon alkioiden mukaan. Tähän ongelmaan ratkaisuksi sopii ristiinvalidointi.

Ristiinvalidointi (cross-validation) on tarkkuuden arvioinnin tapa, jossa data jaetaan aluksi N:een saman kokoiseen osajoukkoon. Tämän jälkeen luodaan N erilaista mallia, käyttäen testijoukkona vuorollaan jokaista N:sta osajoukosta, ja opetusjoukkona koko muuta dataa. Jokaiselle mallille lasketaan yllä kuvattu testijoukon tarkkuus oikein luokiteltujen alkioiden osuutena. Lopuksi eri mallien tarkkuuksista lasketaan keskiarvo kuvaamaan mallin kyvykkyyttä [Kotsiantis, 2007]. Ristiinvalidoinnin avulla satunnaisuuden vaikutusta virheen estimointiin voidaan vähentää merkittävästi.

5.5 Epätasaisten frekvenssien hallinta

Epätasaisista opetusdatan kategoriamäärästä aiheutuu useita huomioonotettavia seikkoja: ennustustarkkuus ei sovi ainoaksi mallin suoriutumisen mittariksi, ja toisaalta taas malli oppii paremmin yliedustetun luokan jakauman. Epätasaisen kategoriajakauman vaikutuksia on yritetty korjata koneoppimisessa aiemmin ainakin kahdella tavalla. Ensimmäinen tapa on tuottaa lisää alkioita vähemmistökategoriaan, tai vähentää alkioita enemmistökategoriasta. Toinen tapa on asettaa *rangaistuksia* (cost) opetusvaiheeseen eri opetusdatan osille [Chawla et al., 2002].

Chawlan, Bowyerin, Hallin ja Kegelmeyerin mukaan vähemmistökategoriaan alkioiden lisäämisen ja enemmistökategoriasta alkioiden vähentämisen yhdistelmä tuottaa paremman lopputuloksen kuin vain toinen näistä yksinään [Chawla et al., 2002]. Heidän vähemmistöalkioiden lisäämiseen kehittämänsä SMOTE-menetelmä käyttää hyväksi piirreavaruutta luoden synteettisiä alkioita, jotka sisältävät ainoastaan piirteiden arvot, eivät itse alkuperäisiä tekstejä. Synteettisiä alkioita ei näinollen voida enää muuttaa alkuperäisiksi

artikkeleiksi [Chawla et al., 2002].

SMOTE:ssa synteettiset alkiot luodaan ottamalla ensin jokaiselle opetusdatan alkiolle a, k lähintä piirreavaruuden naapuria $n_1 ... n_k$, k:n riippuessa tarvittavasta synteettisten alkioiden määrästä. Tämän jälkeen jokaiselle k naapurille arvotaan vektori *gap*, joka sisältää piirteiden määrän satunnaislukuja väliltä [0, 1]. Uuden synteettisen alkion piirrevektorin arvoksi tulee tällöin $gap * (a - n_i)$ [Chawla et al., 2002]. Synteettisten alkioiden avulla päätösrajaa eri kategorioiden välillä voidaan muuttaa yleistettävämmäksi.

Opetusvaiheeseen asetettavat rangaistukset sopivat erityisesti tilanteisiin, joissa erilaiset virheet luokittelussa ovat vähemmän hyväksyttäviä kuin toiset [Shalev-Shwartz ja Ben-David, 2014, s. 232]. Rangaistukset sopivat siis esimerkiksi tilanteisiin, joissa opetusdatassa on yhtä luokkaa enemmän kuin muita, ja luokittimen halutaan suoriutuvan hyvin myös vähemmistöluokkien luokittelusta.

Rangaistuksen asettaminen tapahtuu yleisessä tapauksessa määrittelemällä ensin *häviöfunktio* (loss function) $\Delta : Y \times Y \to \mathbb{R}_{\geq 0}$. Häviöfunktiossa jokaiselle mahdolliselle luokkaparille (y', y) määritellään häviö $\Delta(y', y)$, joka kuvaa virhettä kun luokan y alkio luokitellaan virheellisesti luokkaan y'. Termiä häviöfunktio käytetään myös luokitteluongelmissa joissa erillisiä rangaistuksia ei käytetä, tällöin häviöfunktion arvo on sama kaikille vääriille luokituksille (tällaista tilannetta käsiteltiin luvussa 5.1.1). Ongelma voidaan siis formalisoida funktion $\hat{f}$ optimoinniksi, kun virhettä L halutaan minimoida [Shalev-Shwartz ja Ben-David, 2014, s. 232]:

$$L = \frac{1}{n} \sum_{i=1}^{n} \Delta(\hat{f}(x_i), y_i) \tag{10}$$

Opetusvaiheen lisäksi epätasainen luokkajakauma voidaan ottaa huomioon mallin suoriutumiskyvyn arvioinnissa. Mikäli suoriutumisen mittarina käytetään vain luokittelutarkkuutta, voidaan päätyä tilanteeseen jossa luokittelutarkkuus on suuri, koska yksi luokka dominoi koko opetusdataa. Eräs tämän ongelman huomioonottava mittari on F1-pisteytys [Chinchor, 1992]. F1-pisteytys lasketaan harmonisena keskiarvona tarkkuudesta (precision) ja saannista (recall) seuraavasti [Shalev-Shwartz ja Ben-David, 2014, s. 244]:

$$F1 = \frac{2}{\frac{1}{tarkkuus} + \frac{1}{saanti}} \tag{11}$$

Tarkkuus ja saanti lasketaan jokaiselle luokalle erikseen seuraavasti:

- tarkkuus (precision) = oikeiden ennustusten osuus kaikista luokkaan luokitelluista
- saanti (recall) = oikeiden ennustusten osuus kaikista luokan todellisista alkioista

Tarkkuuden ja saannin ottaminen huomioon samanaikaisesti tekee mallille mahdottomaksi arvata aina opetusdatan yleisintä luokkaa ja saavuttaa korkea pisteytys. Monen luokan luokitteluongelmassa F1-pisteytys lasketaan yleensä keskiarvona kaikkien luokkien F1-pisteytyksistä [Ghamrawi ja McCallum, 2005]. Keskiarvo voi olla painottamaton luokkien välillä, tai esimerkiksi painotettu eri luokkien osuuksilla alkuperäisessä luokitellussa datassa.

5.6 Tutkielmassa käytetyt luokittelijat

5.6.1 Mallin valinnan periaatteet

Kotsiantisin [Kotsiantis, 2007] mukaan itse mallin valinta on kriittinen vaihe mallinnuksessa. Kun malli on valittu ja arvioitu huolellisesti, sitä voidaan opetuksen jälkeen käyttää rutiininomaisesti uuden datan luokittelussa. Mallin valinta koostuu yksinkertaistetusti itse koneoppimismallin valinnasta, parametrien valinnasta, sekä mallin arvioinnista (kuva 5). Kun optimaaliset parametrit tietylle koneoppimismallille on löydetty, palataan koneoppimismallin valintaan ja sen jälkeen taas parametrien valintaan kyseiselle mallille (kuva 5). Mallin ja malliperheen valinta tähtää yleensä yleistysvirheen pienentämiseen.

Mallin valinnassa keskeinen ongelma on, miten monia erilaisia vaihtoehtoja eri kuvan 5 vaiheista tulisi kokeilla parhaan mallin rakentamiseksi [Olson et al., 2016]. Mallin valinnassa voidaan myös priorisoida useita eri tekijöitä, kuten ennustustarkkuutta, mallin tulkittavuutta [James et al., 2014, s. 26] tai suorituskykyä. Usein kun mallin tulkittavuus kasvaa, sen joustavuus vähenee [James et al., 2014, s. 25-26]. Esimerkiksi lineaariset mallit ovat hyviä tulkittavuudeltaan, mutta rajoittuvat vain lineaaristen funktioiden avulla datan kuvaamiseen.

Kokeneilla datatieteilijöillä on usein jonkinlainen käsitys hyvistä malleista ja parametreista tietylle ongelma-alueelle, mutta kokemattomat datatieteilijät saattavat helposti päätyä käyttämään lukemattomasti aikaa erilaisten konfiguraatioiden kokeilemiseen [Olson et al., 2016]. Olsonin, Bartleyn, Urbanowiczin ja Mooren mukaan teoriassa manuaalinen mallin valinta ei oikeastaan ole enää tarpeellistakaan, sillä evoluutiota jäljittelevät algoritmit suoriutuvat nykyaikana hämmästyttävän hyvin tämänkaltaisista ongelmista [Olson et al., 2016].

Kaksi yleistä, moniin ongelmiin sovellettua koneoppimismallia ovat satunnaismetsäluokittelija (random forest classifier) ja tukivektorikone (support vector machine). Seuraavaksi esitellään näiden koneoppimismallien perusperiaatteet.

5.6.2 Satunnaismetsäluokittelija

Satunnaismetsäluokittelija on algoritmi, joka perustuu päätöspuiden käyttöön [Ho, 1995]. Päätöspuu on koneoppimisalgoritmi, jonka muodostamisessa

ja käytössä on tiivistetysti kaksi vaihetta [James et al., 2014, s.306]:

- Jaetaan piirreavaruus, eli piirteiden $X_1, X_2, ..., X_p$ arvot J:een erilliseen alueeseen $R_1, R_2, ..., R_J$. Alueisiin jakamisessa on useita eri lähestymistapoja, mutta pääperiaatteena on parantaa ennustustarkkuutta ja monissa algoritmeissa estää ylisovittamista. Ennustustarkkuuden minimoinnissa yhden alueen jako kahteen alueeseen voidaan formalisoida seuraavan lausekkeen minimoinniksi:

$$\sum_{x_i \in R_1} (y_i - \hat{y}_{R1})^2 + \sum_{x_i \in R_2} (y_i - \hat{y}_{R2})^2 \tag{12}$$

- Jokaista aluetta $R_1, R_2, ..., R_J$ vastaa yksi vastemuuttujan arvo, joka on yleensä keskiarvo opetusdatan kyseiselle alueelle kuuluvien havaintojen vastemuuttujien arvoista (edellisen lausekkeen $\hat{y}_{R1}$ ja $\hat{y}_{R2}$). Jokaiselle uudelle havainnolle palautetaan sitä vastaavan alueen vastemuuttujaarvo.

Päätöspuita voidaan käyttää sekä luokittelu- että regressio-ongelmiin. Satunnaismetsäluokittelija on päätöspuun laajennus kahdella ominaisuudella: useiden puiden käyttämisellä sekä satunnaisuuden lisäämisellä puiden muodostamisprosessiin [Shalev-Shwartz ja Ben-David, 2014, s. 255].

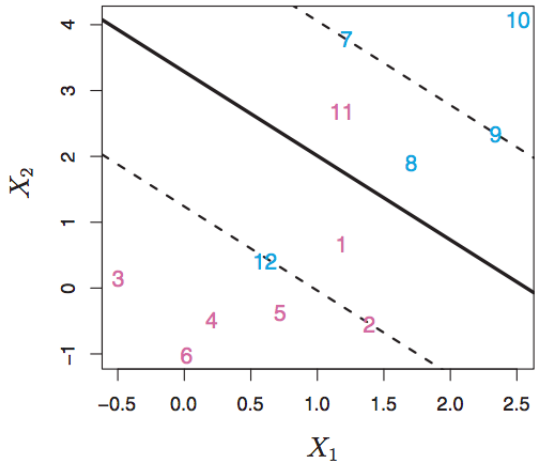
Satunnaismetsäluokittelija koostuu useasta päätöspuusta, joille kaikille muodostetaan oma opetusdatasetti S' satunnaisotannalla takaisinpanolla alkuperäisestä opetusdatasetistä S. Tämän jälkeen jokainen puu muodostetaan siten, että alueita jakaessa jokaisessa halkaisussa otetaan huomioon vain satunnainen osajoukko piirteiden arvoja. Kun kaikki puut on muodostettu, uuden havainnon ennustus lasketaan enemmistöäänestyksellä kaikkien puiden ennustuksista [Shalev-Shwartz ja Ben-David, 2014, s. 255]. Satunnaisten piirteiden käyttäminen alueisiin jakamisessa vähentää eri puiden samankaltaisuutta, mikäli datassa on yksittäisiä piirteitä jotka ovat *vahvoja ennustajia* [James et al., 2014, s.319-320].

Yksi satunnaismetsäluokittelijan hyvä puoli on, että sen tärkeimpinä luokittelussa pitämät piirteet voidaan selvittää kohtalaisen helposti, vaikka useita puita käyttävän luokittelijan selitettävyys onkin huonompi kuin yksittäisen puun [James et al., 2014, s.319]. Yhdelle puulle jokaisen piirteen merkitsevyys voidaan laskea tarkastelemalla niitä puun opetuksen jakoja, jossa on otettu huomioon kyseinen piirre. Yhden piirteen merkitsevyys yhdelle päätöspuulle saadaan selville yhteenlaskemalla näistä jaoista kaavan 11 mukainen neliöidyn virheen pienentyminen samalle alkuperäiselle alueelle jaon jälkeen. Lopulta yhden piirteen merkitsevyys koko satunnaismetsäluokittelijalle voidaan laskea keskiarvona sen merkitsevyydestä jokaiselle puulle. Näitä merkittävyyslukuja vertailemalla voidaan kertoa luokittelijan tärkeimpinä pitämät piirteet [James et al., 2014, s.319].

5.6.3 Tukivektorikone

Tässä aliluvussa esitellään ensin tukivektoriluokittelija [Vapnik ja Lerner, 1963], ja tämän jälkeen sen laajennus tukivektorikone [Cortes ja Vapnik, 1995]. Tukivektoriluokittelija on lineaarinen luokittelija, joka sopii tilanteisiin joissa etsitään luokkia erottelevaa hypertasoa (hyperplane), mutta data ei ole lineaarisesti eroteltavissa [James et al., 2014, s. 344-345].

Tukivektoriluokittelija perustuu hyperpinnan sekä sen ympärillä olevan *marginaalin* optimointiin [James et al., 2014, s. 346]. Tavoitteena on löytää hypertaso sekä sen ympärillä olevien marginaalien leveys, joka mahdollistaa eri luokkien datapisteiden erottelun marginaalien ulkopuolelle, ja minimoi marginaalien sisään jäävien, "väärällä" puolella olevien datapisteiden määrää.



Kuva 6: Tukivektoriluokittelijan havainnollistus kaksiulotteiselle datalle [James et al., 2014, s. 346]

Tukivektoriluokittelija ratkaisee formaalisti seuraavan ongelman:

$$\begin{aligned} &\text{Maksimoidaan M} \\ &(\text{optimoidaan parametreja } M, \beta_0, \beta_1, ...\beta_p, \epsilon_1, ..., \epsilon_n) \\ &\text{edellyttäen} \quad \sum_{j=1}^{p} (\beta_j)^2 = 1 \\ &y_i(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p) \geq M(1 - \epsilon_i) \\ &\epsilon_i \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C, \end{aligned} \tag{13}$$

missä $\beta_0, \beta_1, ...\beta_p$ ovat hyperpinnan parametreja, C positiivinen sovitusparametri (kutsutaan myös rangaistusparametriksi), ja $\epsilon_0, \epsilon_1, ...\epsilon_n$ muuttujia

jotka kertovat missä i:s havainto sijaitsee marginaaleihin ja hypertasoon nähden. Kyseisten muuttujien arvo määritellään seuraavalla tavalla havainnon sijainnin mukaan [James et al., 2014, s. 346-347]:

- Havainnon sijaitessa oikealla puolella marginaalia, $\epsilon_i = 0$
- Jos havainto sijaitsee marginaalin sisällä hypertason oikealla puolella, $\epsilon_i > 0$
- Jos havainto sijaitsee väärällä puolella hypertasoa, $\epsilon_i > 1$

Havaintojen sijaintia kuvaavien muuttujien avulla voidaan myös ymmärtää paremmin sovitusparametrin C merkitystä. Epäyhtälöstä $\sum_{i=1}^{n} \epsilon_i \leq C$ huomataan, että mitä pienemmäksi C on asetettu, sitä vähemmän väärälle puolelle hypertasoa tai marginaalin sisään oikealle puolelle hypertasoa sijoittuvia datapisteitä sallitaan. Parametri C siis vaikuttaa marginaalin leveyteen siten, että C:tä pienennettäessä myös lopputuloksena oleva marginaali pienenee [James et al., 2014, s. 347].

Tukivektorikone on tukivektoriluokittelijan laajennus, joka mahdollistaa tukivektoriluokittelijan käyttämisen myös tilanteessa, jossa luokkien välisen rajan halutaan olevan epälineaarinen [James et al., 2014, s. 349]. Tukivektorikoneen intuitio perustuu havaintojen muuntamiseen piirreavaruutta useampiulotteisempaan avaruuteen, jossa lineaarinen hypertaso erottelee datapisteet paremmin [Shalev-Shwartz ja Ben-David, 2014, s. 215-216]. Lopputuloksena on epälineaarinen luokitteluraja alkuperäisessä piirreavaruudessa.

6 Ohjatun koneoppimisen sovellus kehysten tunnistamiseen

Tässä luvussa kuvaillaan johdannossa esitettyihin tutkimuskysymyksiin vastaamiseksi rakennettu koneoppimismalli, jonka rakennuksessa on hyödynnetty kaikissa edellisissä luvuissa käsiteltyä teoriaa. Ensimmäisessä aliluvussa kuvaillaan datankeruu sekä opetusdatan tuottaminen tutkimuskysymykseen sopivilla menetelmillä. Toisessa aliluvussa listataan datasta eristetyt piirteet. Kolmannessa aliluvussa käsitellään datan käsittelyssä tehtyjä valintoja, ja neljännessä aliluvussa perustellaan työssä käytettyjen luokittelijoiden valinta.

6.1 Datankeruu ja esivalmistelut

6.1.1 Datankeruu

Tutkimuskysymyksiin vastaamiseen käytettiin dataa suomalaisesta verkkomediana toimivasta vastamediasta MV-lehdestä. MV-lehti voidaan kirjoitushetkellä katsoa Suomen suosituimmaksi vastamediaksi [Ylä-Anttila, 2018],

jonka poliittisia linjoja voidaan kuvailla esimerkiksi maahanmuuttovastaisiksi ja oikeistolaisiksi.

Valitsin MV-lehden tutkimuksen datalähteeksi, koska sen ympärillä on käyty voimakasta yhteiskunnallista keskustelua, ja siitä löytyi myös tarpeeksi dataa laskennallisen analyysin tarpeisiin. Hain dataa myös viidestä muusta suomalaisesta vastamediasta, joista MV-lehdestä löytyi määrällisesti eniten artikkeleita.

Datankeruu tutkimukseen tehtiin MV-lehden verkkosivuilta verkkosivujen haravoinnilla (web scraping) [Sanastokeskus TSK, 2013] Python-ohjelmointikielellä. Yhteensä MV-lehdestä saatiin 37 477 artikkelia, jotka tarkastettiin ohjelmallisesti uniikeiksi. Jokaisesta artikkelista kerättiin HTML-sisältö sekä artikkelin linkki.

Taulukko 1: Artikkelien määrät julkaisuvuosittain

2014	2015	2016	2017	1–2/2018	Yhteensä
372	9,947	15,266	10,875	1,017	37,477

Tutkimuskysymyksiin vastaamiseksi datasta tuli eritellä ne uutiset, joissa viitataan yhteen tai useampaan valtamedian uutiseen. Tätä varten jokaisesta kerätystä artikkelista tuli eristää sen sisältämät lähdelinkit. Tein lähdelinkkien eristämisen säännöllisten lausekkeiden avulla siten, että loin 25 kiellettävän merkkijonon listan, jotka sisältävät linkit suodatettiin pois. Loin listan laadullisen artikkelien tarkastelun perusteella siten, että otin listalle yleisimmät selkeät mainostamiseen tai sosiaalisen median jakamiseen liittyvät merkkijonot. Kiellettyjä merkkijonoja olivat esimerkiksi ”twitter.com/share” sekä ”.flexcard.fi”.

Artikkelin linkkien domainien eristämisen jälkeen tuli selvittää, mitkä viitedomaineista ovat valtamediasivustoja. Sitä varten hyödynsin olemassaolevaa 81 valtamedian listaa. Lista oltiin tuotettu aiemmin Nelimarkan, Laaksosen ja Semaanin [Nelimarkka et al., 2018] luokittelukehikon mukaan. Listan tuottamisessa oltiin käytetty 6 suomalaisen vastamedian artikkeliaineiston eniten viitatuimpia lähdedomaineja. Listan medioita olivat esimerkiksi ”iltalehti.fi”, ”hs.fi” ja ”telegraph.co.uk”.

Suodatin kaikista artikkeleista valtamediaviitteen sisältävät artikkelit ottamalla mukaan ne artikkelit, jotka sisälsivät viitteen johonkin edellisessä kappaleessa kuvaillun 81 median listan domaineista. Yhteensä näitä artikkeleita tuli 10906.

6.1.2 Kehystämisen prosessien identifiointi

Kun olin suodattanut datasta valtamedialinkin sisältävät uutiset, vuorossa oli artikkelien laadullinen analyysi ensimmäiseen tutkimuskysymykseen vastaamiseksi. Tavoitteenani oli identifioida tapoja, joilla MV-lehden artikkeleissa

uudelleenkehystetään valtamedian uutisia, sekä tapoja joilla artikkeleissa kehystetään itse valtamediaa siihen viittaamisen yhteydessä.

Kuten luvussa 4.1 todettiin, monet kehysanalyysiä soveltavat tutkijat ovat noudattaneet vain löyhästi muiden tutkijoiden kehittämiä kehysanalyysin määritelmiä. Oma tutkimukseni perustuu Entmanin luvussa 4.1 esitettyyn määritelmään: kehystäminen on ”joidenkin näkökantojen valitsemista ja tekemistä merkittävämmäksi tekstillä kommunikoiden” [Entman, 1993]. Otin myös vaikutteita Gamsonin ja Laschin luvussa 4.1 esitellystä *allekirjoitusmatriisista*.

Laadullinen kehysanalyysini perustui pääosin MV-lehden verkkolehden lukemiseen. Saadakseni mahdollisimman hyvän kuvan myös lehden vanhemmista artikkeleista, otin 84 artikkelin satunnaisotoksen koko datasetin niistä artikkeleista, jotka sisälsivät jonkun lähdelinkin, vuosilta 2015-2018. Tein jokaisesta näistä artikkelista lyhyesti muistiinpanot Gamsonin ja Laschin allekirjoitusmatriisin tyylisesti, mutta kuitenkin kevyemmin siten, että kirjasin lyhyesti artikkelin näkökulman sekä näkökulmalle ominaisia sitaatteja.

Esimerkiksi muistiinpanoista joita kirjasin 84 artikkelista, muutamasta artikkelista olin kirjannut, että kyseessä on rikosuutinen jossa on lähteenä valtamediauutinen, jota siteerataan paljon. Sitaateiksi olin kirjannut esimerkiksi eräästä Li Anderssonia arvostelevasta jutusta rivin ”Imagea tulee ja pitää boikotoida”. Muistiinpanojen tarkoituksena oli hahmottaa yleiskuvaa MV-lehden juttujen aiheista, näkökulmista ja niiden suhteista käytettyihin lähteisiin. Otin mukaan myös ne artikkelit joiden lähteenä ei ollut valtamedia, jotta saisin mahdollisimman kattavan kuvan siitä, miten MV-lehti kehystää erilaisia lähdelinkkejä.

Artikkelimuistiinpanojen, sekä useita kuukausia kestäneen MV-lehden päivittäisen lukemisen avulla hahmottelin kolme tapaa, joilla MV-lehti kehystää valtamedian uutisia tai itse valtamediaa. Kutsun niitä seuraavassa kuvauksessa *kategorioiksi*.

Journalistisen median kritiikki Ensimmäisen kategorian uutisissa kritisoidaan journalistista mediaa. Kritisointi voi olla esimerkiksi boikotointiin kehottamista, median tai sen tekstien arvostelua, tai kriittisten mielipiteiden esittämistä median toiminnasta. Tähän kategoriaan kuuluu myös median epäkohdista, kuten sensuroinnista uutisoiminen.

Esimerkkejä alla:

- *Otsikko*: “Iltalehden Islamia myötäilevä propagandapaljastus-kuva!!”
- *Otsikko*: “virallinen media antaa uutiskatsauksissaan tapahtumista niin höttöisen kuvan, ettei siihen usko edes Pihtiputaan mummo”
- *Otsikko*: “Imagea tulee ja pitää boikotoida”

Sisällön kopiointi valtamediasta: Toisen kategorian artikkelit kopioivat sisältöä valtamediasta: niillä on sama aihe ja joskus täysin sama tarina

kuin lähdeartikkelilla. Lähdeartikkelista saattaa olla lainattu tekstiä, tai joskus lähdeartikkelin teksti on kopioitu kokonaan. Tekstin suora kopiointi lähdeartikkelista ei kuitenkaan ole edellytys tähän kategoriaan kuulumiseen. Jos lähdeartikkelin tarinaa ei olla kopioitu sellaisenaan, sitä on uudelleenkehystetty lisäämällä sisältöä ohjaamaan lukutapaa. Joissain artikkeleissa argumentointia tukemassa on ylimääräisiä lähteitä, kuten linkkejä hallinnolliseen dataan aiheesta. Esimerkkejä alla:

- *Otsikko*: “Kitee vastasi: Ei tilaa mamuille!! Hyvä Kitee!!’"
  Lainaus: “Kiteen kaupungilta ei löydy tiloja, joihin voisi majoittaa tilapäisesti vähintään sata maahanmuuttajaa . . . ” (lause kopioitu valtamedialähdeartikkelista)
  “Tämä on kaupunginhallituksen vastaus ely-keskukselle, joka kartoittaa vastaanottokeskukseksi sopivia tiloja Suomen kunnista. ”
- *Otsikko*: “Keskusta vahvistaa: Tavoitteena III-oluen poisto kaupoista!!’"
  Lainaus: “Keskusta haluaa poistaa keskioluen ruokakaupoista ja palauttaa sen myynnin takaisin Alkoon, selviää puolueen julkistamasta vaaliohjelmasta. Puoluesihteeri Timo Laaninen vahvistaa asian Uudelle Suomelle ja toteaa . . . ” (lause kopioitu valtamedialähdeartikkelista) ”

Oman narratiivin rakentaminen lähdeviitteiden avulla: Tämän kategorian artikkeleissa on eri aihe ja narratiivi kuin lähdeartikkelissa. Artikkeleissa käytetään lähdeviitteitä tukemaan tarinaa. Artikkelin aihe saattaa kuitenkin olla kopioitu esimerkiksi sosiaalisen mediasta tai blogipostauksesta, jossa viitataan valtamediauutiseen. Esimerkkejä alla:

- *Otsikko*: “Pimittääkö Fimea toimivia syöpä- ja epilepsiahoitoja suomalaisilta?”
  “. . . Lisää lähteitä jutun painoksi:
  CNN-kanavan ohjelma liittyen aiheeseen [linkki CNN:n sivustolle]”
- *Artikkeli* (pitkästä artikkelista, jonka aiheena on, onko NATO uhka Suomen turvallisuudelle):
  “Kuten New Orleansin mafiapomo Carlos Marcellon työhuoneessa olevassa plakaatissa sanottiin: [linkki *Washington Postiin*]”
  Lainaus: “Kolme voi pitää salaisuuden, jos kaksi on kuollut.” (teksti otettu *Washington Postin* artikkelista)
- *Otsikko*: “Hankamäki: Vasemmistolaista monikulttuuri-ideologiaa yliopistolla’"
  Lainaus: “Mainitsin parin viikon takaisessa kirjoituksessani siitä,

että yliopiston ja aikuisten maailman väliin perustetusta Tiedekulmasta on tulossa kovaa vauhtia tieteen ja tutkimuksen popularisointia varten rakennettu esiintymislava . . . ”

“Viihteestä pitivät huolta minulle ennestään tuntematon Ailu Valle ja Mannerheimia sotarikollisena pitävä [linkki *Ilta-Sanomiin*] rap-artisti Paleface . . . ”

Alussa olin jakanut toisen kategorian, *Sisällön kopioinnin valtamediasta*, kahdeksi kategoriaksi, eli kategorioita oli yhteensä siis neljä. Alkuperäiset kahden eri sisällön kopioinnin kategorian nimet olivat *Lähdejutun uudelleenkehystäminen valikoinnilla tai kopioinnilla* sekä *Lähdejutun uudelleenkehystäminen arvottavalla puheella*.

Ensimmäiseen edellämainituista kategorioista kuuluivat uutiset, joissa lähdejutusta oltiin neutraalisti valikoitu ja kopioitu tekstiä sekä tarinan osasia, ilman arvottavaa kommentointia. Arvottavalla kommentoinnilla tarkoitetaan tässä tutkielmassa kielenkäyttöä, jolla alkuperäisen lähdeartikkelin tapahtumista tai henkilöistä luodaan negatiivisempi tai positiivisempi kuva kuin lähdeartikkelissa. Toiseen kategoriaan kuuluivat artikkelit, joissa oltiin valikoitu ja kopioitu sisältöä lähteestä kommentoiden samalla alkuperäistä juttua arvottavasti. Esimerkkejä näistä kahdesta kategoriasta ovat yhdistetyn kategorian esimerkit: ensimmäinen esimerkki Kiteen toiminnasta kuului arvottavalla puheella kommentointiin, sillä heti sen otsikossa tokaistiin: ”Hyvä Kitee!!”, joka voidaan katsoa positiivisesti arvottavaksi puheeksi. Seuraava, III-oluen kaupoista poistoon liittyvä esimerkki taas raportoi neutraalisti Keskustan kannanotosta.

Sisällön kopioinnin kategorioiden yhdistämisen syynä oli, että arvottavaa puhetta oli hyvin hankala erottaa neutraalista puheesta objektiivisesti. Neljän kategorian hahmottelemisen jälkeen tein kollegani kanssa testin, jossa molemmat luokittelimme itsenäisesti samat 50 MV-lehden artikkelia neljään kategoriaan. Keskustelimme tuloksista ja huomasimme, että monen artikkelin kohdalla meillä oli eriävä mielipide siitä, kumpaan sisällönkopiointikategoriaan artikkeli kuuluu. Esimerkiksi eräässä artikkelissa pohdittiin jihadistien radikalisointikoulutuksesta ”Onkohan Suomessa jo vastaavaa toimintaa?”, jonka arvottavuudesta olimme aluksi eri mieltä. Totesin, että mikäli ihminenkään ei osaa erottaa näitä kahta kategoriaa toisistaan, olisi luultavasti hyvin vaikeaa rakentaa automaattista luokittelijaa näille kategorioille.

6.1.3 Opetusdatan luokittelu

Kun opetusdatan luokituskategoriat olivat selvillä, teimme luokittelun objektiivisuuden luotettavuutta mittaavan Kappa-testin [Cohen, 1960]. Saavutimme testissä Kappa-arvon 0.59. Eri aloilla on erilaiset standardit hyvälle kappa-arvolle: esimerkiksi biostatistiikan julkaisussa [McHugh, 2012] 0.59 katsotaan heikoksi arvoksi. Kuitenkin yhteiskuntatieteellisessä mediatutki-

muksessa 0.59 voidaan katsoa hyväksi arvoksi: esimerkiksi Weberin [Weber, 2014] julkaistussa tutkimuksessa raportoitiin useita alle 0.5 Kappa-arvoja. Weberin uutisista tunnistamia ominaisuuksia olivat esimerkiksi *valta*, *yksilöllisyys* ja *ristiriita*, joiden kaikkien Kappa-arvot olivat alle 0.5.

Varsinainen opetusdata tuotettiin ottamalla 1000 artikkelin satunnaisotos luvussa 6.1.1 mainitusta 10906 artikkelin datasetistä, joista jokainen uutinen sisälsi linkin valtamediaan. Tämän jälkeen luokittelin jokaisen uutisista yhteen neljästä luokasta:

1. Journalistisen median kritiikki
2. Sisällön kopiointi valtamediasta
3. Oman narratiivin rakentaminen lähdeviitteiden avulla
4. Ei mahdollista luokitella

Koska jokainen luokista 2-4 voi sisältää myös journalistisen median kritiikkiä, päätin, että luokka 1 on dominoiva: mikäli artikkeli sisältää journalistisen median kritiikkiä, se luokitellaan luokkaan 1, riippumatta siitä sisältääkö se myös piirteitä muista luokista. Tietojenkäsittelytieteen alalla tunnetaan myös menetelmiä moneen luokkaan yhtäaikaiseen luokitteluun (*multi-label classifying*) [Boutell et al., 2004, Ghamrawi ja McCallum, 2005]. Eräs mahdollinen mallinnustapa olisikin ollut rakentaa monen luokan luokittelija tälle ongelmalle. Päädyin kuitenkin poissulkevaan yhteen luokkaan luokittelijaan, koska hyvin pienessä osassa uutisia tapahtui journalistisen median kritiikin ja muiden luokkien uutisten limittymistä.

Taulukko 2: Luokitellun opetusdatan kategorioiden frekvenssit

Luokka 1	Luokka 2	Luokka 3	Luokka 4	Yhteensä
72	711	122	95	1000

Luokitellun opetusdatan osuudet olivat epätasaisesti jakautuneet (taulukko 2). Luokkien epätasaisuutta on mahdollista hallita useilla eri tavoilla, joita käsiteltiin luvussa 5.5.

Luvussa 6.1.2 esitellyssä eri kategorioiden kuvauksissa ei otettu kantaa siihen, kuinka monta valtamedialähdeartikkelia tulee olla, ja miten luokittelu tehdään jos valtamedialähdeartikkeleita on useampi. Suurimmassa osassa luokiteltavia artikkeleita luokittelu oli mahdollista tehdä ilman tätä jakoa: monissa artikkeleissa oltiin selkeästi joko referoitu pääosin yhtä valtamedia-artikkelia (luokka 2) tai käytetty monipuolisesti argumentoinnin tukena yhtä tai useampaa lähdeartikkelia (luokka 3).

Useimmissa artikkeleissa, joissa luokittelua ei ollut mahdollista tehdä (luokka 4), oltiin joko referoitu kahta valtamedia-artikkelia ja muodostettu

näistä jonkinlainen sekoitus, tai referoitu pääosin yhtä artikkelia, mutta sekoitettu mukaan niin paljon omaa argumentointia, että uutisen luokitteleminen johonkin näistä kategorioista tuntui mahdottomalta. Luokittelin tällaiset artikkelit omaan kategoriaansa ”ei voi luokitella”.

6.2 Piirteiden eristäminen

Opetusdatan luokittelun jälkeen eristin datasta piirteitä ohjattua luokittelijaa varten. Piirteiden eristämistä varten tuotin datasta alkuperäisen HTML-artikkelidatan lisäksi kaksi eri versiota: tekstiversion, sekä karsitun (stemmed) version. Tekstiversiossa oltiin poistettu kaikki muut merkit paitsi numerot ja sanat, karsitussa versiossa taas muutettu tekstimuodon sanat perusmuotoon [Lovins, 1968]. Molemmista muodoista oltiin myös poistettu 847 sanan lista pysäytyssanoja (stop words).

Esiprosessoinnin jälkeen eristin datasta seuraavat piirteet:

Sanaston ominaisuudet

- TF-IDF (term frequency–inverse document frequency) [Jones, 1972, Robertson, 2004] karsitulle artikkelille.
- Karsitun artikkelin pituus merkeissä
- Erikoismerkkien käyttö alkuperäisessä artikkelissa: merkkien ”!”, ”?”, ”?!”, ”...” määrät.
- Joidenkin aineiston kannalta merkittävien sanojen määrät alkuperäisessä artikkelissa: esimerkiksi lehtien nimiä (HS, hesari, iltalehti) sekä sanojen ”toimittaja” ja ”sensuuri” muunnoksia. Eristin tämän ominaisuuden sen vuoksi, että karsintakirjastojen (stemming) tunnetaan toimivan vaihtelevasti suomen kielelle. Koska en ollut varma mihin muotoon karsinta muuntaa sanoja, halusin varmistaa näiden sanojen luotettavamman piirteiksi muuntamisen.

HTML-tagien määrät artikkelissa

Seuraavien tagien määrät alkuperäisessä artikkelissa:

- Otsikot *h2* ja *h3*
- *Iframe*
- *Strong*
- *Blockquote*
- *a* (linkki)

HTML-tagien sisällä olevat tekstit

HTML-tagien ominaisuudet eristettiin jokaiselle tagille erikseen alkuperäisestä artikkelista. HTML-tagien sisällä olevista teksteistä eristettiin seuraavat ominaisuudet tageille *blockquote* ja *strong*:

- Artikkelin kaikkien kyseisen tagin sisällä olevien tekstien yhteispituus jaettuna artikkelin pituudella
- Pisimmän tagin sisällä olevan tekstin pituus
- Keskimääräinen tagin sisällä oleva tekstipituus

HTML-tagien sijoittuminen artikkeliin

Seuraavat ominaisuudet eristettiin tageille *a* (linkki) ja *iframe* (sisällön upotus) alkuperäisestä artikkelista:

- Keskimääräinen etäisyys kahden tagin välillä
- Pisin pituus kahden tagin välillä
- Lyhin pituus kahden tagin välillä

Kuvien ominaisuudet

Seuraavat ominaisuudet eristettiin *img* -tagille (kuva) alkuperäisestä artikkelista

- Etäisyys ensimmäiseen kuvaan
- Keskimääräinen kuvakoko leveys*korkeus -muodossa
- Suurin kuvakoko leveys*korkeus -muodossa
- Kuvatekstien erisnimien (isolla kirjoitettujen sanojen, jotka eivät ole lauseen alussa) yhteislukumäärä
- Kuvatekstien yhteispituus artikkelissa

Tekstin tyyliominaisuudet

Seuraavat ominaisuudet eristettiin artikkelille tekstiversiosta:

- Isolla alkukirjaimella kirjoitettujen sanojen määrä koko artikkelitekstistä
- Caps lockilla kirjoitettujen sanojen määrä koko artikkelitekstistä

Lähdeviitteiden käyttö

Jokaisesta aineistossa esiintyvästä lähdeviitedomainista muodostettiin oma piirteensä siten, että jos domain esiintyi artikkelissa enemmän kuin kerran, piirteen arvoksi koodattiin 1, muuten 0. Päätin rajata piirteen arvon yhteen, sillä joissakin artikkeleissa huomasin olevan epäluonnollisen paljon yhden domainin linkkejä, ja arvelin niiden voivan muodostua virhelähteeksi.

6.3 Valinnat datan käsittelyssä

Koska omassa työssäni dataan kuului vain 1000 artikkelia, päätin valita opetusjoukon koon jakaen datasta 90% opetusjoukkoon ja 10% testijoukkoon. Alkioiden valinnan tein satunnaisesti, käyttäen ristiinvalidointia tarkkuuden arvioinnin luotettavuuden parantamiseksi.

Alkuperäisessä datan luokittelussa käytin luokkaa 4, ”ei mahdollista luokitella”. Alustavan data-analyysin jälkeen huomasin kuitenkin, että kyseisen luokan artikkelit eivät ole yhtenäinen luokka, vaan joukko artikkeleita jotka muistuttavat jokainen jonkun kolmen muun luokan edustajaa. Päätin luokitella ”ei voi luokitella” -kategorian alkiot kolmeen muuhun kategoriaan sen mukaan, minkä kategorian edustajia ne muistuttavat eniten luvun 6.1.2 kehysten prosessien kuvausten mukaisesti.

Luokka 1	Luokka 2	Luokka 3
82	770	148

Taulukko 3: Luokitellun datan frekvenssit ”ei mahdollista luokitella” luokan poistamisen jälkeen

”Ei mahdollista luokitella” -luokan alkioiden uudelleenluokittelun jälkeen tuloksena oli edelleen epätasainen luokkien jakauma (taulukko 3). Käytin luvussa 5.5 kuvattua SMOTE-menetelmää luokkien epätasaisuuden hallintaan. SMOTE-menetelmää ja ristiinvalidointia käyttäen algoritmini toimi seuraavalla tavalla:

1. Sekoitetaan luokiteltu data satunnaisesti järjestettyyn listaan pareja (x, y), missä x on opetusdatan alkio ja y sitä vastaava luokka.
2. Jaetaan luokiteltu data 10 uniikkiin osaan. Jokaiselle 10 tavalle valita yksi osa testijoukoksi ja muut osat opetusjoukoksi:
   (a) Sovelletaan SMOTE-menetelmää opetusjoukon frekvenssien tasapainottamiseksi
   (b) Opetetaan valittu malli kyseisellä opetusjoukolla
   (c) Tallennetaan mallin tarkkuus oikein luokiteltujen alkioiden osuutena testijoukosta
3. Annetaan mallin lopulliseksi tarkkuudeksi keskiarvo eri jaoilla tallennetuista tarkkuuksista

6.4 Malliperheen valinta

Valitsin työhöni kaksi mallia, satunnaismetsäluokittelijan ja lineaarisen tukivektorikoneen. Valitsin satunnaismetsäluokittelijan sen vuoksi, että se on

Satunnaismetsäluokittelija	PU	PI	MIN	Tarkkuus	F1
1	10	sqrt(N)	2	0.789	0.735
2	10000	sqrt(N)	2	0.804	0.745
3	10	N	2	0.773	0.765
4	10	sqrt(N)	10	0.805	0.773
5	1000	1000	2	0.813	0.774

Taulukko 4: Yleistysvirhe satunnaismetsäluokittelijoille. PU= puiden määrä, PI = huomioon otettavien piirteiden määrä, N = piirteiden kokonaismäärä, MIN = vähimmäismäärä alkioita jotka vaaditaan puun jakamisessa sisäisen solmun jakoon. F1-pisteytys luokkien osuuksien opetusdatassa mukaan painotettu keskiarvo.

joustavampi menetelmä kuin lineaariset luokittelijat, mutta kuitenkin perusperiaatteiltaan paremmin tulkittavissa oleva kuin esimerkiksi epälineaariset tukivektorikoneet [James et al., 2014, s. 25]. Lineaarisen tukivektorikoneen valitsin sen vuoksi, että halusin selvittää, kuinka hyvin yli 30 000 piirrettä sisältävä luokiteltu data on eroteltavissa lineaarisesti.

7 Kokeet

Valitsin vertailuun kaksitoista erilaista mallia: neljä satunnaismetsäluokittelijaa ja kahdeksan tukivektorikonetta. Satunnaismetsäluokittelijat valitsin vaihtelemalla kolmea parametria: puiden määrää, puun jakamisessa huomioonotettavien piirteiden määrää sekä pienintä määrää alkioita jotka vaaditaan puun jakamisessa sisäisen solmun jakoon. Tukivektorikoneet valitsin vaihtelemalla rangaistusparametria.

7.1 Mallien yleistysvirheet

Taulukoissa 4 ja 5 listataan eri mallien vaihtelevien parametrien arvot, sekä mallien luokittelutarkkuudet edellisessä luvussa kuvaillulla ristiinvalidoinnilla. Tuloksista huomataan, että satunnaismetsäluokittelijat suoriutuvat luokittelusta huomattavasti paremmin kuin tukivektorikoneet. Tästä voidaan päätellä, että data ei ole lineaarisesti eroteltavissa. Taulukon 6 sekaannusmatriisissa (confusion matrix) havainnollistetaan parhaiten suoriutuneen satunnaismetsäluokittelijan tekemiä luokituksia eri alkuperäisten luokkien alkioille.

Luokittelija	C	Tarkkuus	F1
1	0.001	0.325	0.335
2	0.01	0.385	0.414
3	0.1	0.369	0.384
4	1	0.369	0.384
5	10	0.369	0.384
6	100	0.369	0.384
7	1000	0.369	0.384
8	10000	0.369	0.384

Taulukko 5: Tukivektorikoneen yleistysvirhe eri rangaistusparametreilla. C = rangaistusparametri, F1-pisteytys luokkien osuuksien opetusdatassa mukaan painotettu keskiarvo.

	Luokka 1	Luokka 2	Luokka 3
Luokka 1	1	4	2
Luokka 2	0	79	2
Luokka 3	0	5	7

Taulukko 6: Sekaannusmatriisi (confusion matrix) eräälle satunnaismetsäluokittelijalle (taulukon 4 luokittelija 5 eräällä ristiinvalidoinnin jaolla opetus- ja testidatajoukkoon). Sekaannusmatriisin i:nnen rivin j:nnen kolumnin arvo kuvaa sitä määrää artikkeleita, jotka on luokiteltu luokkaan i mutta kuuluvat luokkaan j.

7.2 Luokittelussa merkitsevimmät piirteet

Eri luokittelijoiden opettamisen jälkeen selvitin, mitkä piirteet merkitsevät luokittelussa eniten. Valitsin tarkasteltavaksi luokittelijaksi satunnaismetsäluokittelija numero 5:n (taulukko 4), koska se saavutti pienimmän yleistysvirheen. Otin tarkasteluun kyseisen mallin samalla opetus-testidatajaolla kuin taulukossa 6 sekaannusmatriisissa käytetty jako. Kaksikymmentä merkitsevintä piirrettä tärkeysjärjestyksessä ovat listattuna taulukossa 6.

	Piirre		Piirre
1	toimittaj (TF-IDF)	11	yle (TF-IDF)
2	iltasanom (TF-IDF)	12	media (TF-IDF)
3	jutu (TF-IDF)	13	mis (TF-IDF)
4	toimittaja (mainintamäärä)	14	pisin pituus kahden linkin välillä
5	stemmatun artikkelin pituus	15	yle (TF-IDF)
6	kirjoit (TF-IDF)	16	<h3> (mainintamäärä)
7	? (mainintamäärä)	17	kuvakaappaus (TF-IDF)
8	seurauks (TF-IDF)	18	medioiden mainintamäärät
9	tämä (TF-IDF)	19	lähd (TF-IDF)
10	suome (TF-IDF)	20	uutis (TF-IDF)

Taulukko 7: Satunnaismetsäluokittelijan 20 merkitsevintä piirrettä tärkeysjärjestyksessä.

Taulukosta 7 huomataan, että lähes kaikista merkitsevimmistä TF-IDF -piirteistä johdettavat sanat liittyvät mediaympäristöön. Esimerkiksi kaksi merkittävintä piirrettä ovat ”toimittaj” ja ”kirjoit”, ja kahdeksanneksi merkittävin piirre on ”iltasanom”, josta voidaan suoraan johtaa suomalaisen Iltasanomat -lehden nimi. Tulos antaa näyttöä siitä, että kehystämisen prosessien tunnistamista voidaan tehdä laskennallisesti, vaikka itse artikkelit kertovatkin eri aiheista: olennaista on, että jokaisessa aineiston artikkelissa hyödynnetään eri tavalla valtamediaa.

Toinen mielenkiintoinen havainto merkittävimmistä piirteistä on, että myös muilla piirteillä kuin käytetyillä sanoilla tai sanajuurilla (TF-IDF) on merkitystä luokittelussa. Usein ohjatuissa luokittelijoissa käytetään vain sanoihin liittyviä piirteitä, mutta tämän luokittelijan perusteella myös artikkelin muotoiluun ja rakenteeseen liittyviä piirteitä kannattaa mahdollisuuksien mukaan eristää datasta ja käyttää luokittelussa. Taulukosta 7 huomataan, että esimerkiksi stemmatun artikkelin pituus sekä otsikkojen käyttö (<h3>)

ovat merkittävimpien piirteiden joukossa.

7.3 Merkitsevimpien piirteiden jakaumat eri luokkien artikkeleissa

Merkittävimpien piirteiden identifioimisen jälkeen selvitin, millä tavalla näiden piirteiden arvot eroavat toisistaan eri luokissa. Laskin jokaiselle 20:stä merkitsevimmästä piirteestä keskiarvot jokaisen eri luokan artikkelien suhteen.

Taulukosta 8 huomataan, että merkitsevimpien piirteiden keskiarvot eri luokkien suhteen tuovat uutta tietoa kehystämisen prosesseista. Esimerkiksi, sanoja ”toimittaj”, ”kirjoit”, ”iltasanom”, sekä ”media” käytetään enemmän journalistisen median kritiikin luokassa kuin muissa luokissa. Toisaalta taas keskimäärin pisimmät artikkelit sanamäärissä ovat oman narratiivin rakennuksen luokassa.

8 Keskustelu

Tässä luvussa esitellään tutkielman rajoitteita, sen merkitystä eri tieteenaloilla sekä jatkotutkimusaiheita.

8.1 Rajoitteet

Laskennallisten menetelmien soveltaminen yhteiskuntatieteisiin sisältää useita erilaisia päätöksiä, ja näistä seuraa tutkimukselle erilaisia rajoitteita. Salganik [Salganik, 2017] on listannut laskennallisille yhteiskuntatieteille tyypillisiä tutkimusrajoitteita, jotka liittyvät niin datan laatuun, teoreettisten konseptien operationalisointiin, kuin analyysimenetelmiinkin. Seuraavaksi esitellään joitakin tämän tutkimuksen rajoitteita Salganikin esimerkkien avulla.

Operationalisoinnilla tarkoitetaan teoreettisten konseptien ilmaisua epätäydellisen datan avulla [Salganik, 2017], ja Salganikin mukaan tämä on usein unohdettu vaihe rajoitteiden tarkastelussa. Tässä tutkielmassa operationalisointi tapahtui tutkimuskysymysten valinnassa sekä päätöksessä tunnistaa kehystämisen prosesseja laskennallisesti ohjatulla luokittelijalla. Koska tässä tutkielmassa ei operationalisoitu esimerkiksi eri kehystämisen prosesseja suoraan tietyksi ominaisuuksiksi datassa vaan luokiteltiin opetusdata laadullisesti, operationalisointi on huomattavasti pienempi rajoite kuin tutkimuksissa, joissa teoreettiset konseptit esitetään suoraan datan ominaisuuksina.

Toinen Salganikin mainitsema yleinen rajoite on datan epäedustavuus. Vaikka yhteiskuntatieteissä usein tavoitellaankin datan edustavuutta, täysin edustavan datan hankinta on vaikeaa datatieteissä [Salganik, 2017]. Myös tässä tutkielmassa datan epäedustavuus on rajoite: datasta voi puuttua joitakin

Piirre	Kritiikki	Kopiointi	Oma narratiivi	Keskiarvo
toimittaj (TF-IDF)	0.02413	0.0022	0.00678	0.00472
iltasanom (TF-IDF)	0.0221	0.0022	0.00028	0.00357
jutu (TF-IDF)	0.03059	0.00402	0.01166	0.00738
toimittaja (mainintamäärä)	0.00016	2e-05	4e-05	4e-05
stemmatun artikkelin pituus	1975.0	1324.05951	3125.17647	1650.47333
kirjoit (TF-IDF)	0.01375	0.00388	0.01248	0.00601
? (mainintamäärä)	0.00067	0.00043	0.00059	0.00048
seurauks (TF-IDF)	0.00064	0.00134	0.00525	0.00187
tämä (TF-IDF)	0.01017	0.0062	0.0127	0.00751
suome (TF-IDF)	0.02713	0.02587	0.04364	0.02866
yle (TF-IDF)	0.02141	0.00438	0.00424	0.00578
media (TF-IDF)	0.02185	0.003	0.01003	0.00564
mis (TF-IDF)	0.00808	0.00219	0.00251	0.00273
pisin pituus kahden linkin välillä	947.46667	535.88099	1329.21324	690.06111
yle (TF-IDF)	0.03044	0.00707	0.00522	0.00874
<h3> (mainintamäärä)	0.00015	7e-05	0.00013	9e-05
kuvakaappaus (TF-IDF)	0.02931	0.01008	0.00685	0.01119
medioiden mainintamäärät	1.08	0.17126	0.44853	0.28889
lähd (TF-IDF)	0.01044	0.01469	0.00516	0.0129
uutis (TF-IDF)	0.02018	0.00555	0.00647	0.00691

Taulukko 8: Piirteiden arvojen keskiarvot jokaisen eri luokan artikkelien suhteen erään opetus-testidatajaon opetusdatalle. Kritiikki -kolumnilla tarkoitetaan luvussa 6.1.2 esiteltyä ”Journalistisen median kritiikki” -kategoriaa, Kopiointi -kolumnilla kategoriaa ”Sisällön kopiointi valtamediasta” ja Oma narratiivi -kolumnilla kategoriaa ”Oman narratiivin rakentaminen lähdeviitteiden avulla”. Neljännessä kolumnissa ”Keskiarvo” esitetään piirteen arvojen keskiarvo kaikkien tarkastelussa olevien artikkelien suhteen.

datalähteen jo sivustoltaan poistamia vanhoja artikkeleita, ja lisäksi valtamedialinkin sisältävien artikkeleiden rajauksessa kaikkia valtamediadomaineja ei pystytty identifioimaan. Tämä voi aiheuttaa vääristymän esimerkiksi kehysten prosessien identifiointiin. Mikäli datassa on esimerkiksi valtamedialinkin sisältäviä artikkeleita, jotka hyödyntävät täysin erilaista kehystämisen prosessia kuin mitä identifioin tutkielmassani, tällöin identifioimani kehystämisen prosessit eivät kuvaa luotettavasti MV-lehden valtamediauutisten kehystämistä. Tämän tutkielman kontribuutio onkin enemmän menetelmällinen kuin yhteiskuntatieteellinen datalähteen toimintaa selittävä työ.

Kolmas Salganikin mainitsema rajoite on datan ”likaisuus”, tarkoittaen sillä datan sisältöä joka ei ole lainkaan tutkijoiden mielenkiinnon kohteena [Salganik, 2017]. Tässä tutkielmassa ”likaista” sisältöä ovat artikkeleiden mainoslinkit, joita suodatin pois mutta joita jäi edelleen dataan analyysivaiheessa. Mainoslinkit eivät kuitenkaan varsinaisesti vaikuta tutkimukseni tuloksiin, sillä identifioin valtamedialinkkidomainit manuaalisesti. Käytin ohjatussa luokittelijassani piirteinä linkkidomaineja, jotka ovat voineet sisältää mainoslinkkidomaineja, mutta ainakaan ne eivät päätyneet merkittävimmiksi piirteiksi.

Neljäs Salganikin mainitsema rajoite on muutos (drifting) [Salganik, 2017], jota sivuttiin myös tämän tutkielman luvussa 3.1. Hän tarkoittaa sillä yleisesti ottaen tilannetta, jossa laskennallinen operationalisointi ei enää vastaa todellisuutta: esimerkiksi tietyn keskusteluaiheen twiittejä hashtagien avulla seurattaessa tilannetta, jossa hashtageillä löytyvä keskustelu eivät vastaa enää alkuperäistä keskusteluaihetta. Tällöin hashtageilla löytyvästä keskustelusta ei voida enää tehdä päätelmiä alkuperäisestä keskustelunaiheesta [Salganik, 2017]. Tässä tutkielmassa muutos rajoittaa rakentamieni luokittelijoiden käyttöä tulevaisuudessa, sillä datalähteen kirjoitustyyli todennäköisesti muuttuu ajan kuluessa eikä luokittelija enää mittaa samaa asiaa uudelle datalle kuin opetusdatalle.

Salganikin mainitsemien yleisten rajoitteiden lisäksi työllä on ainakin kaksi muuta rajoitetta: laadullisen luokittelun subjektiivisuus sekä mallin tarkkuus sen soveltamisessa päätelmien tekoon. Laadullisen luokittelun subjektiivisuutta sivuttiin luvussa 5.2.1, ja myös tämän tutkielman ongelma on yksittäisten artikkeleiden opetusdataan luokitteluun liittyvä luokittelijan subjektiivisuus. Lopuksi, koska opetusdatassa luokka 2 on selkeästi yliedustettuna, mallin hyödyntäjän tulee kriittisesti arvioida sen tarkkuutta käsillä olevan sovellustehtävään.

8.2 Työn merkitys eri tieteenaloilla

Tämän tutkielman kontribuutioita voidaan esitellä monella tieteen tekemisen abstraktiotasolla. Tässä aliluvussa kerrotaan ensin, miten tutkielma sijoittuu ja kontribuoi yhteiskuntatieteissä laajemmin meneillään olevaan digitaalisten aineistojen ja menetelmien kehitykseen. Sen jälkeen kommentoidaan

tämän työn merkitystä jatkuvassa muutoksessa olevalle tieteenalalle datatieteelle. Lopuksi esitellään yksityiskohtaisemmin juuri tämän tutkielman kontribuutiot laskennalliseen mediatutkimukseen.

Yhteiskuntatieteissä digitaalisten aineistojen käyttöönotto on ollut huomattavasti hitaampaa kuin vaikkapa biologiassa tai fysiikassa, vaikka laskennallisilla menetelmillä on monia mahdollisuuksia yhteiskuntatieteissä [Lazer et al., 2009]. Kun aiemmin datan määrä kasvoi suoraan lineaarisessa suhteessa sen keräämiseen käytettyyn aikaan esimerkiksi haastatteluita tehdessä tai lehtiartikkeleita käsin kerätessä, laskennallisessa datankeruussa kerättävän datan määrää voidaan skaalata helpommin [Albanese, 2010]. Sosiaalisista verkostoista kerättävä data tarjoaa myös uudenlaisia mahdollisuuksia selvittää, mitä ihmiset todella tekevät, kun usein tutkitaan vain ihmisten ajattelua [boyd, 2010, Albanese, 2010].

Boydin [boyd, 2010] mukaan pelkät laskennalliset menetelmät eivät kuitenkaan riitä uudenlaisen yhteiskuntatieteen tekemiseen, sillä tarvitaan myös tutkijoita jotka osaavat tulkita ilmöitä ja esimerkiksi ymmärtävät, että ihmisten toiminta sosiaalisessa mediassa ei ole aina intentionaalista. Hän esittää kaksi mahdollista ratkaisua: tietojenkäsittelytieteilijöiden ja yhteiskuntatieteilijöiden yhteistyö, sekä peruskoulutus myös toiselta alalta niin tietojenkäsittelytieteilijöille kuin yhteiskuntatieteilijöillekin. Tämä tutkielma on tehty ikään kuin näiden kahden ratkaisun yhdistelmänä: tein tutkielman osana poikkitieteellistä tutkimusryhmää ja -projektia, mutta kuitenkin itsenäisesti sekä teoria- että empiriaosuuksia työstäen.

Datatieteelle tieteenalana ei ole olemassa yksikäsitteistä määritelmää, mutta esimerkiksi Van der Aalstin mukaan datatiede on uusi tietojenkäsittelytieteestä, matematiikasta sekä useista muista aloista erilleen erkaantumassa oleva ala, joka vastaa haasteeseen saada hyödyllistä informaatiota kasvavista datamääristä [van der Aalst, 2014]. Hänen mukaansa datatieteen voidaan jaotella vastaavan seuraaviin kysymyksiin datan avulla: mitä tapahtui, miksi tapahtui, mitä tulee tapahtumaan ja mikä olisi parasta mitä voisi tapahtua. Tämä tutkielma on eräs datatieteellinen kontribuutio ja esimerkki siitä, miten usein datatieteilijää tarvitaan joskus myös muotoilemaan ratkaistavan ongelman kysymyksiä, ei vain itse ratkaisuja.

Tämän tutkielman tarkemmat tieteelliset kontribuutiot keskittyvät mediatutkimuksen alalle. Tutkielman luvusta 4 käy ilmi, että laskennallista kehysanalyysiä ei ole tehty uudelleenkehystämiselle, eli tilanteelle jossa kehystetään toisen lähteen tekstiä uuteen merkitykseen. Tässä työssä esitettiin ohjatun koneoppimisen sovellus valtamediauutisten uudelleenkehystämisen analysointiin, joten se kontribuoi laskennallisen kehysanalyysin menetelmälliseen kehitykseen.

Toiseksi, tämä työ kontribuoi suomalaiseen vastamedia- ja valeuutistutkimukseen. Suomalaisten vastamedioiden toiminnasta tiedetään edelleen melko vähän, vaikka aihe on tunnustettu tärkeäksi. Tämä työ identifioi vastamedian tapoja hyödyntää ja rakentaa luottamusta valtamedian avulla.

8.3 Jatkotutkimuskohteet

Luvussa 7.3 esiteltiin merkitsevimpien piirteiden keskiarvoja luokittain. Piirteiden arvot toivat lisää tietoa eri kehystämisen prosesseissa käytetyistä sanoista. Toisaalta taas tiedämme tämän tutkielman luvusta 2, että valeuutisiksi kutsutut artikkelit eivät usein sisällä väärää tietoa. Kun tiedämme nyt, että eri kehystämisen prosesseissa käytetään erilaisia sanoja, ja että suomalaisille valeuutisille on tyypillistä virheellisen tiedon poissaolo, eräs mielenkiintoinen jatkotutkimusaihe on laskennallinen kehysanalyysi hyödyntäen konvoluutioneuroverkkoja. Niiden avulla voitaisiin yrittää poimia erilaisissa vastamediauutisissa käytettyjä kielellisiä keinoja tarkemmin kuin pelkästään sanojen esiintyvyysmääriä yksittäin mittaavilla malleilla.

Toinen mielenkiintoinen jatkotutkimuskohde on tämän tutkielman kehystämisen prosessien kategoria 2, sisällön kopiointi valtamediasta. Luvun 6.1.2 mukaan osassa tämän kategorian artikkeleissa tekstiä oltiin otettu valtamedian lähdeuutisesta melko suoraviivaisesti, kun taas joissakin artikkeleissa alkuperäiseen uutiseen oltiin lisätty erilaista arvottavaa kuvailua. Sisällön kopioinnin tutkimusta voisikin siis jatkaa suoraan selvittämällä laskennallisesti, kuinka paljon ja mihin aiheisiin liittyen sisältöä kopioidaan valtamediasta vastamediaan, sekä kehittämällä menetelmiä kopioinnin yhteydessä käytettyjen argumentaatiokeinojen kuvailuun.

Kolmas jatkotutkimuskohde on luvussa 6.1.3 mainittu moneen luokkaan samanaikaisesti luokittelu. Tässä tutkielmassa huomattiin, että poissulkevien kategorioiden kehittäminen ja niiden pohjalta luokittelijan rakentaminen ei ole yksinkertaista. Luokittelija joka mahdollistaisi moneen luokkaan yhtäaikaisen luokittelun, helpottaisi opetusdatan manuaalista luokittelua ja voisi tarjota tarkempia tuloksia luokittelijan kuvatessa realistisemmin todellisuutta.

9 Yhteenveto

Valeuutiset ovat viime aikoina nousseet merkittäväksi yhteiskunnalliseksi puheenaiheeksi [Allcott ja Gentzkow, 2017], erityisesti vaaleihin liittyvään vaikuttamiseen liittyen. Valeuutisten on arveltu vaikuttaneen merkittävästi vuoden 2016 Yhdysvaltojen presidentinvaalien tulokseen, ja Suomessa taas sosiaalisen median yhtiö Facebook ja Suomen valtion edustajat toimivat yhteistyössä jakaen tietoa Suomen vuoden 2019 eduskuntavaaleihin liittyvästä informaatiovaikuttamisesta [Mansikka, 2019].

Vaikka mediassa usein puhutaan valeuutisista yhtenäisenä käsitteenä, tutkimuksessa valeuutiselle ei ole yhtenäistä ja kattavaa määritelmää. Esimerkiksi Allcott and Gentzkow määrittelevät valeuutiset “uutisartikkeleiksi, jotka ovat tarkoituksellisesti and todistettavasti valhetta, ja voivat harhaanjohtaa lukijaa” [Allcott ja Gentzkow, 2017]. Laskennallinen valeuutimus liittyy lähinnä niiden lukijakunnan tutkimukseen sekä valeuutisten tunnista-

miseen. Aiemmasta tutkimuksesta Yhdysvalloissa tiedetään esimerkiksi, että valeuutissivustot saavat enemmän vierailuja Facebookista kuin valtamediasivustot [Nelson ja Taneja, 2018]. Valeuutisten tunnistamisen algoritmiset lähestymistavat voidaan jakaa kahteen ryhmään: valeuutisten tunnistamiseen lingvististen ominaisuuksien perusteella [Pérez-Rosas et al., 2017] sekä artikkeleiden jakoverkoston perusteella [Kumar ja Geethakumari, 2014].

Suomessa monet valeuutiset eivät sisällä virheellistä tietoa, joten niitä kutsutaan tutkimuksessa myös vastamedioiksi [Noppari ja Hiltunen, 2017]. Aiemmasta tutkimuksesta tiedetään, että suomalaiset vastamediauutiset hyödyntävät valtamedian uutisia tukemaan omia agendojaan uudelleenkehystämällä niitä [Ylä-Anttila, 2018, Noppari ja Hiltunen, 2017]. Tästä seuraa tutkielmani motivaatio selvittää tarkemmin, miten vastamediauutisissa kehystetään valtamedian uutisia, ja miten kehystämisen prosessien identifioinnin voisi automatisoida.

Viestinnän tutkimuksessa kehystämisellä tarkoitetaan prosessia, jossa valintojen ja jäsentämisen avulla muokataan mediaesityksen tulkintaa [Seppänen, 2014, s. 97], ja kehystämisen erilaisia määritelmiä käsitellään tarkemmin tutkielman luvussa neljä. Kehysanalyysiin ei ole vakiintunutta menetelmää datatieteessä, mutta sitä on tehty sekä ohjatuilla [Burscher et al., 2014] että ohjaamattomilla [Pashakhin, 2016] menetelmillä.

Tämän tutkielman tutkimuskysymyksinä on selvittää, miten suomalaisissa vastamediauutisissa kehystetään valtamedian uutisia, ja millaisella menetelmällä kehystämisen prosessien tunnistaminen voidaan automatisoida. Tutkielman empiirinen osuus sisälsi datankeruun, kehysten identifioinnin sekä ohjatun koneoppimismenetelmän kehittämisen vaiheet. Data kerättiin suomalaisesta vastamediasta MV-lehdestä, ja siitä suodatettiin noin 11 000 valtamedialinkin sisältävää artikkelia.

Aineistosta identifioitiin kolme kehystämisen prosessia, joihin luokiteltiin satunnaisotos valtamedialinkin sisältävistä artikkeleista. Nämä kolme kehystämisen prosessia ovat journalistisen median kritiikki, sisällön suora tai epäsuora kopiointi journalistisesta mediasta sekä oman narratiivin argumentointi valtamediaviitteiden avulla. Eri kehystämisen prosessien identifioinnin jälkeen luokitellusta datasta eristettiin piirteitä, joita käytettiin erilaisten satunnaisluokittelijoiden sekä tukivektorikoneiden opetuksessa.

Tulosten mukaan parhaiten tutkielman luokitteluongelmaan sopii satunnaismetsäluokittelija, ja lineaarisen tukivektorikoneen huonosta suoriutumisesta voidaan päätellä, että piirredata ei ole lineaarisesti eroteltavissa. Eräs mielenkiintoinen tulos liittyy satunnaismetsäluokittelijan tärkeimpinä pitämiin piirteisiin: suuri osa sanoina tai TF-IDF-juurina käytetyistä piirteistä olivat mediasanastoa. Tästä voidaan päätellä, että vastamedian erilaisia orientaatioita lähdesivustoon voidaan tunnistaa laskennallisesti lähteeseen liittyvän sanaston käytön avulla.

Tärkeimpien piirteiden joukossa oli myös artikkelin HTML-elementteihin liittyviä ominaisuuksia, joita käytetään yleensä luokittelupiirteinä vähemmän

kuin sanoja tai sanajuuria. Kun tiedetään että vasta- ja valemediat käyttävät usein omalaatuisia tyylillisiä HTML-ratkaisuja, voidaan tästä päätellä, että HTML-elementteihin liittyviä ominaisuuksia kannattaa käyttää valeuutistutkimuksessa. Yhteenvetona, vaikka luokittelijaa ei voida pitää tarpeeksi tarkkana useimpiin käytännön korkeaa tarkkuutta vaativiin sovelluksiin, siitä saatiin kuitenkin uutta kuvailevaa tietoa eri kehystämisen prosesseissa käytetystä sanastosta ja tyylillisistä keinoista.

Lähteet

[Albanese, 2010] Albanese, E. (2010). Scaling Social Science with Apache Hadoop . `https://blog.cloudera.com/blog/2010/04/scaling-social-science-with-hadoop/`. Luettu: 2019-03-23.

[Allcott ja Gentzkow, 2017] Allcott, H. ja Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2).

[Blei et al., 2003] Blei, D. M., Ng, A. Y., ja Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022.

[Boutell et al., 2004] Boutell, M. R., Luo, J., Shen, X., ja Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition*, 37:1757–1771.

[boyd, 2010] boyd, d. (2010). Big Data: Opportunities for Computational and Social Sciences. `http://www.zephoria.org/thoughts/archives/2010/04/17/big-data-opportunities-for-computational-and-social-sciences.html`. Luettu: 2019-03-23.

[Breiman, 2001] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

[Burscher et al., 2014] Burscher, B., Odijk, D., Vliegenthart, R., de Rijke, M., ja de Vreese, C. H. (2014). Teaching the Computer to Code Frames in News: Comparing Two Supervised Machine Learning Approaches to Frame Analysis. *Communication Methods and Measures*, 8(3):190–206.

[Burscher et al., 2016] Burscher, B., Vliegenthart, R., ja Vreese, C. H. (2016). Frames Beyond Words: Applying Cluster and Sentiment Analysis to News Coverage of the Nuclear Power Issue. *Social Science Computer Review*, 34(5):530–545.

[Carragee ja Roefs, 2004] Carragee, K. M. ja Roefs, W. (2004). The neglect of power in recent framing research. *Journal of Communication*, 54(2):214–233.

[Chang et al., 2009] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., ja Blei, D. M. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. Teoksessa Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., ja Culotta, A., toimittajat, *Advances in Neural Information Processing Systems 22*, sivut 288–296. Curran Associates, Inc.

[Chawla et al., 2002] Chawla, N. V., Bowyer, K. W., Hall, L. O., ja Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence research*, 16(1):321–357.

[Chinchor, 1992] Chinchor, N. (1992). Muc-4 evaluation metrics. Teoksessa *Proceedings of the 4th Conference on Message Understanding*, MUC4 '92, sivut 22–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Cohen, 1960] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

[Cortes ja Vapnik, 1995] Cortes, C. ja Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[D'Angelo, 2002] D'Angelo, P. (2002). News Framing as a Multiparadigmatic Research Program : A Response to Entman. (December):870–888.

[David et al., 2011] David, C. C., Atun, J. M., Fille, E., ja Monterola, C. (2011). Finding Frames: Comparing Two Methods of Frame Analysis. *Communication Methods and Measures*, 5(4):329–351.

[Day ja Thompson, 2012] Day, A. ja Thompson, E. (2012). Live From New York, It's the Fake News! Saturday Night Live and the (Non)Politics of Parody. *Popular Communication*, 10(1-2):170–182.

[Debye, 1909] Debye, P. (1909). Näherungsformeln für die Zylinderfunktionen für große Werte des Arguments und unbeschränkt veränderliche Werte des Index. *Mathematische Annalen*, 67(4):535–558.

[Edell, 2018] Edell, A. (2018). I trained fake news detection AI with >95 accuracy, and almost went crazy. `https://towardsdatascience.com/i-trained-fake-news-detection-ai-with-95-accuracy-and-almost-went-crazy-d10589aa57c`. Luettu: 2019-03-22.

[Entman, 1993] Entman, R. M. (1993). Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication*, 43(4):51–58.

[European commission, 2018] European commission (2018). *A multi-dimensional approach to disinformation.*

[Gamson ja Lasch, 1983] Gamson, W. A. ja Lasch, K. E. (1983). The Political Culture of Social Welfare Policy. *Evaluating the Welfare State Social and Political Perspectives*, 95(221):397–415.

[Geman ja Geman, 1984] Geman, S. ja Geman, D. (1984). Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741.

[Ghamrawi ja McCallum, 2005] Ghamrawi, N. ja McCallum, A. (2005). Collective multi-label classification. *Proceedings of the 14th ACM international conference on Information and knowledge management - CIKM '05*, sivu 195.

[Goffman, 1974] Goffman, E. (1974). *Frame analysis: An Essay on the Organization of Experience*.

[Golbraikh ja Tropsha, 2000] Golbraikh, A. ja Tropsha, A. (2000). Predictive QSAR modeling based on diversity sampling of experimental datasets for the training and test set selection. *Molecular Diversity*, 5(4):231–243.

[Goldberg, 2017] Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, 10:1–309.

[Goodfellow et al., 2016] Goodfellow, I., Bengio, Y., ja Courville, A. (2016). *Deep Learning*. MIT Press. `http://www.deeplearningbook.org`.

[Greene et al., 2014] Greene, D., O'Callaghan, D., ja Cunningham, P. (2014). How Many Topics? Stability Analysis for Topic Models. Teoksessa Calders, T., Esposito, F., Hüllermeier, E., ja Meo, R., toimittajat, *Machine Learning and Knowledge Discovery in Databases*, sivut 498–513, Berlin, Heidelberg. Springer Berlin Heidelberg.

[Greussing ja Boomgaarden, 2017] Greussing, E. ja Boomgaarden, H. G. (2017). Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies*, 43(11):1749–1774.

[Hellsten et al., 2010] Hellsten, I., Dawson, J., ja Leydesdorff, L. (2010). Implicit media frames: Automated analysis of public debate on artificial sweeteners. *Public Understanding of Science*, 19(5):590–608.

[Ho, 1995] Ho, T. K. (1995). Random decision forests. Teoksessa *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, sivut 278–, Washington, DC, USA. IEEE Computer Society.

[Horsti, 2005] Horsti, K. (2005). *Vierauden Rajat*. Tampere University Press.

[Hosseinimotlagh ja Papalexakis, 2018] Hosseinimotlagh, S. ja Papalexakis, E. E. (2018). Unsupervised Content-Based Identification of Fake News Articles with Tensor Decomposition Ensembles.

[Hu et al., 2014] Hu, B., Lu, Z., Li, H., ja Chen, Q. (2014). Convolutional Neural Network Architectures for Matching Natural Language Sentences. Teoksessa Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., ja Weinberger, K. Q., toimittajat, *Advances in Neural Information Processing Systems 27*, sivut 2042–2050. Curran Associates, Inc.

[James et al., 2014] James, G., Witten, D., Hastie, T., ja Tibshirani, R. (2014). *An Introduction to Statistical Learning: With Applications in R*. Springer Publishing Company, Incorporated.

[Jones, 1972] Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.

[Kannan ja Gurusamy, 2015] Kannan, S. ja Gurusamy, V. (2015). Preprocessing Techniques for Text Mining. (October 2014).

[Koltsov et al., 2014] Koltsov, S., Koltsova, O., ja Nikolenko, S. I. (2014). Latent Dirichlet Allocation: Stability and Applications to Studies of User-generated Content. *Proceedings of the 2014 ACM Conference on Web Science*, sivut 161–165.

[Kotsiantis, 2007] Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. Teoksessa *Proceedings of the 2007 Conference on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies*, sivut 3–24, Amsterdam, The Netherlands, The Netherlands. IOS Press.

[Kumar ja Geethakumari, 2014] Kumar, K. P. K. ja Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-centric Computing and Information Sciences*, 4(1):14.

[Lang, 1995] Lang, K. (1995). Newsweeder: Learning to filter netnews. Teoksessa *in Proceedings of the 12th International Machine Learning Conference (ML95*.

[Lazer et al., 2009] Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., ja Van Alstyne, M. (2009). Computational Social Science. *Science*, 323(5915):721–723.

[Lease, 2011] Lease, M. (2011). On quality control and machine learning in crowdsourcing. *AAAI Workshop - Technical Report*, WS-11-11:97–102.

[Lovins, 1968] Lovins, J. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(June):22–31.

[Mansikka, 2019] Mansikka, O. (2019). Suomi haluaa välttää trump-vaalien kohtalon - valtio ja facebook sopivat niistä keinoista, joilla vaaleihin varaudutaan. `https://www.hs.fi/nyt/art-2000005964858.html`. Luettu: 2010-03-21.

[Marchi, 2012] Marchi, R. (2012). With Facebook, blogs, and fake news, teens reject journalistic 'objectivity'. *Journal of Communication Inquiry*, 36(3):246–262.

[Matthes ja Kohring, 2008] Matthes, J. ja Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2):258–279.

[McHugh, 2012] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, sivut 276–282.

[Miller, 1997] Miller, M. M. (1997). Frame Mapping and Analysis of News Coverage of Contentious Issues. *Social Science Computer Review*, 15(4):367–378.

[Nelimarkka et al., 2018] Nelimarkka, M., Laaksonen, S.-M., ja Semaan, B. (2018). Social Media Is Polarized, Social Media Is Polarized. *Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18*, sivut 957–970.

[Nelson ja Taneja, 2018] Nelson, J. L. ja Taneja, H. (2018). The small, disloyal fake news audience: The role of audience availability in fake news consumption. *New Media & Society*, 20(10):3720–3737.

[Noppari ja Hiltunen, 2017] Noppari, E. ja Hiltunen, I. (2017). Only an idiot would search for objective truth" populist counter media from the perspective of audience studies. *Media Point Conference, Prague, Czech Republic*.

[Olson et al., 2016] Olson, R. S., Bartley, N., Urbanowicz, R. J., ja Moore, J. H. (2016). Evaluation of a tree-based pipeline optimization tool for automating data science. Teoksessa *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, GECCO '16, sivut 485–492, New York, NY, USA. ACM.

[Olver et al., 2019] Olver, F. W. J., Daalhuis, A. B. O., Lozier, D., Schneider, B. I., Boisvert, R. F., Clark, C. W., Miller, B. R., ja Saunders, B. V. (2019). http://dlmf.nist.gov/, Release 1.0.22 of 2019-03-15.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., ja Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

[Pashakhin, 2016] Pashakhin, S. (2016). Topic modeling for frame analysis of news media. Teoksessa *Proceedings of the Artificial Intelligence and Natural Language AINL FRUCT 2016 Conference, Saint-Petersburg, Russia, 10-12 November 2016*, sivut 103–106.

[Pérez-Rosas et al., 2017] Pérez-Rosas, V., Kleinberg, B., Lefevre, A., ja Mihalcea, R. (2017). Automatic Detection of Fake News. *arXiv e-prints*, sivu arXiv:1708.07104.

[Porter, 1997] Porter, M. F. (1997). Readings in information retrieval. kappale An Algorithm for Suffix Stripping, sivut 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Qazvinian et al., 2011] Qazvinian, V., Rosengren, E., Radev, D. R., ja Mei, Q. (2011). Rumor has it: Identifying misinformation in microblogs. Teoksessa *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, sivut 1589–1599, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Quinn ja Bederson, 2011] Quinn, A. J. ja Bederson, B. B. (2011). Human computation: A survey and taxonomy of a growing field. Teoksessa *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, sivut 1403–1412, New York, NY, USA. ACM.

[Robertson, 2004] Robertson, S. (2004). Understanding inverse document frequency: On theoretical arguments for IDF. *Journal of Documentation*, 60(5):503–520.

[Roh et al., 2018] Roh, Y., Heo, G., ja Euijong Whang, S. (2018). A Survey on Data Collection for Machine Learning: a Big Data - AI Integration Perspective. *arXiv e-prints*, sivu arXiv:1811.03402.

[Salganik, 2017] Salganik, M. J. (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, Princeton, NJ, open review edition.

[Sanastokeskus TSK, 2013] Sanastokeskus TSK (2013). Tiedonharavointi. `http://www.tsk.fi/tsk/fi/haku-266.html?page=get_id&id=ID331&vocabulary_code=TSKTT`. Luettu: 2019-03-21.

[Seppänen, 2014] Seppänen, J. (2014). *Mediayhteiskunta*. Vastapaino, Tampere.

[Settles ja Craven, 2008] Settles, B. ja Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. Teoksessa *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, sivut 1070–1079, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Shalev-Shwartz ja Ben-David, 2014] Shalev-Shwartz, S. ja Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA.

[Shu et al., 2019] Shu, K., Wang, S., ja Liu, H. (2019). Beyond news contents: The role of social context for fake news detection. Teoksessa *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, sivut 312–320, New York, NY, USA. ACM.

[Silverman, 2016] Silverman, C. (2016). This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. `https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook?utm_term=.icMK1vp5X#.bg62MBwYQ`. Luettu: 2019-03-21.

[Subramanian, 2017] Subramanian, S. (2017). Inside the macedonian fake-news complex. `https://www.wired.com/2017/02/veles-macedonia-fake-news/`. Luettu: 2019-03-21.

[van der Aalst, 2014] van der Aalst, W. M. P. (2014). Data scientist : The engineer of the future.

[Vapnik ja Lerner, 1963] Vapnik, V. ja Lerner, A. (1963). Pattern Recognition using Generalized Portrait Method. *Automation and Remote Control*, 24:774–780.

[Väliverronen, 1996] Väliverronen, E. (1996). *Ympäristöuhkan anatomia : tiede, mediat ja metsän sairaskertomus*. Vastapaino, Tampere.

[Wang et al., 2012] Wang, D., Kaplan, L., Le, H., ja Abdelzaher, T. (2012). On truth discovery in social sensing: A maximum likelihood estimation approach. Teoksessa *Proceedings of the 11th International Conference on Information Processing in Sensor Networks*, IPSN '12, sivut 233–244, New York, NY, USA. ACM.

[Wardle ja Derakhshan, 2017] Wardle, C. ja Derakhshan, H. (2017). Information Disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*.

[Weber, 2014] Weber, P. (2014). Discussions in the comments section: Factors influencing participation and interactivity in online newspapers' reader comments. *New Media and Society*, 16(6):941–957.

[Wilbur ja Sirotkin, 1992] Wilbur, W. J. ja Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science*, 18(1):45–55.

[Wu et al., 2014] Wu, L., Morstatter, F., Hu, X., ja Liu, H. (2014). Mining Misinformation in Social Media. *Big Data in Complex and Social Networks*, sivut 1–34.

[Yin et al., 2008] Yin, X., Han, J., ja Yu, P. S. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, 20(6):796–808.

[Yin ja Tan, 2011] Yin, X. ja Tan, W. (2011). Semi-supervised truth discovery. Teoksessa *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, sivut 217–226, New York, NY, USA. ACM.

[Ylä-Anttila, 2018] Ylä-Anttila, T. (2018). Populist knowledge: 'Post-truth' repertoires of contesting epistemic authorities. *European Journal of Cultural and Political Sociology*, (January):1–33.